PRODUCTIVITY IN MORPHOLOGICAL NEGATION:
A CORPUS-BASED APPROACH

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Humanities

2010

VICTORIA HULSE

SCHOOL OF EDUCATION
List of Contents

List of Tables and Figures 6
Abstract 8
Declaration 9
Copyright Statement 10
Acknowledgements 11

Introduction 12

Chapter 1 Research Context 16

1.1 Introduction 17

1.2 Productivity 19
1.2.1 The scope of productivity 22
1.2.2 The importance of productivity in language 24
1.2.3 Areas of relevance to productivity 24

1.3 Negation 28
1.3.1 What is negation? 30
1.3.2 Morphological negation as a case study 31

1.4 Corpus Linguistics 33
1.4.1 Attitudes to corpus-based approaches 34
1.4.2 The rationale for using a corpus 37

1.5 Research context summary 40

Chapter 2 Literature Review 42

2.1 Early studies in Productivity - intuitive and dictionary-based approaches 44

2.2 Corpus-based approaches 49

2.3 Psycholinguistic approaches 60

2.4 Combination approaches 65

2.5 Literature summary 69
Chapter 3  Data Collection: developing the database and examining the prefixes

3.1  Why a corpus-based approach?  72
   3.1.1  Why the British National Corpus?  76

3.2  Which negative prefixes and why?  76

3.3  Creating a database of negatively-prefixed words  81
   3.3.1  Use of software to identify negatively-prefixed words  82
   3.3.2  Editing the database  83
   3.3.3  Accuracy of the BNC and Sketch Engine word-class tagging  87
   3.3.4  The final database  87

3.4  Historical factors affecting the negative prefixes  89
   3.4.1  Small historical ‘snap-shot’ study  93

3.5  Semantic factors affecting the negative prefixes  95
   3.5.1  *A*-prefix semantics  95
   3.5.2  *Dis*-prefix semantics  97
   3.5.3  *In*-prefix semantics  98
   3.5.4  *Non*-prefix semantics  101
   3.5.5  *Un*-prefix semantics  105
   3.5.6  Semantic factors summary  107

3.6  Data collection and prefix research summary  108

Chapter 4  Analysis and Discussion

4.1  Replicating methods for measuring productivity  114
   4.1.1  The ‘type’ method  116
      4.1.1.1  Aronoff’s (1976) type/potential type method  116
      4.1.1.2  Bolozky’s (1993; 1999) dictionary method  118
      4.1.1.3  Type counts for negative prefixes  120
      4.1.1.4  Discussion of the ‘type’ method  122
   4.1.2  The ‘token’ method  124
      4.1.2.1  Discussion of the ‘token’ method  129
   4.1.3  The ‘type/token’ method  130
      4.1.3.1  Discussion of the ‘type/token’ method – Measure A  131
   4.1.4  The ‘hapax/token’ method  132
      4.1.4.1  Discussion of the ‘hapax/token’ method – Measure B  133
   4.1.5  The ‘hapax/type’ method  135
      4.1.5.1  Discussion of the ‘hapax/type’ method – Measure C  137
      4.1.5.2  Overview of the Five Measurements  138
4.1.5.3 Frequencies of low-occurring words: the hapax legomena, dislegomena and trislegomena 139
4.1.5.4 Analysis of type frequencies over greater frequency-distribution range 141
4.1.5.5 Comparison of type frequencies over the 1-3 range and the 1-20 range 144
4.1.6 A critique of Measures A, B and C 145

4.2 Examining the hapaxes 150
4.2.1 Identifying the correlation between hapaxes and neologisms 155
4.2.2 The application of disqualification criteria to determine ‘genuine’ neologisms 157
4.2.3 To what extent do hapaxes reflect morphological productivity? 159
4.2.4 Proposed new methods: the ‘neologism/type’ method and the ‘neologism/token’ method 161

4.3 The role of word class in morphological productivity 163
4.3.1 Word class and type distribution 165
  4.3.1.1 Do negatively-prefixed types favour adjectives, adverbs, nouns or verbs? 165
  4.3.1.2 Do types belonging to each word class favour different prefixes? 168
  4.3.1.3 Negatively-prefixed types compared with the whole BNC according to word class 168
4.3.2 Word class and token distribution 170
  4.3.2.1 Do negatively prefixed tokens favour adjectives, adverbs, nouns or verbs? 170
  4.3.2.2 Do tokens belonging to each word class favour different prefixes? 173
  4.3.2.3 Negatively-prefixed tokens compared with the whole BNC according to word class 174
4.3.3 Summary of the role of word class in morphological productivity 175
4.3.4 Two case studies: word class types and tokens for specific prefixes 176
  4.3.4.1 A- prefixed adjectival types and the hapax/type ratio 177
  4.3.4.2 Dis- prefixed verbal tokens and the hapax/token ratio 178
4.3.5 Measure A – Word class productivity – mean type frequency for each prefix 180
4.3.6 Applying the neologism/token (Measure B) and neologism/type (Measure C) ratios to the word class distributions 183

4.4 Which negative prefix favours which domain? 186

4.5 Chapter summary 189
# Chapter 5 Conclusion

## 5.1 Introduction

## 5.2 Research findings

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.1</td>
<td>Which measures of productivity have been replicated?</td>
</tr>
<tr>
<td>5.2.2</td>
<td>How useful are hapax legomena in measuring productivity?</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Which measure of productivity is most appropriate?</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Which prefix is the most productive?</td>
</tr>
<tr>
<td>5.2.5</td>
<td>How is productivity affected by word class?</td>
</tr>
<tr>
<td>5.2.6</td>
<td>Historical factors and their impact on the productivity of the negative prefixes</td>
</tr>
<tr>
<td>5.2.7</td>
<td>Semantic factors and their impact on the productivity of the negative prefixes</td>
</tr>
</tbody>
</table>

## 5.3 Have the aims of this research been fulfilled?

## 5.4 Limitations, recommendations and future research

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.1</td>
<td>Limitations of this research</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Recommendations and future research</td>
</tr>
</tbody>
</table>

## 5.5 Chapter and thesis summary

## Bibliography

(Database disc – inside back cover)

Word count: 58,470
List of Tables and Figures

Chapter 2

Table 2.1 Summary Matrix of Negative Prefixes after Funk (1971:374) 47

Chapter 3

Table 3.1 Comprehensive list of potential negative prefixes 79
Table 3.2 Negatively-prefixed types in the BNC – including breakdown into word classes 88
Table 3.3 Frequencies of negatively-prefixed tokens in the BNC – including breakdown into word classes 88

Figure 3.1 Introduction of the positive and the negated forms of the highest-frequency negatively-prefixed words from the database 94
Table 3.4 Summarising the semantic tendencies of the negative prefixes (after Hamawand, 2009:131-137) 107

Chapter 4

Table 4.1 Types (in 1976), Additions (by 1996), and proportion of new types (after Hulse, 2002, based on Bolozky, 1999) 119
Table 4.2 Negative Prefixes – Total number of negatively-prefixed types in the BNC 121
Table 4.3 Negative Prefixes – Total number of tokens for database types 127
Table 4.4 Negative Prefixes – types, tokens and type/token figures 131
Table 4.5 Negative Prefixes – hapaxes, tokens and hapax/token figures in the BNC 133

Figure 4.1 Hapax/token ratio (after Baayen and Lieber’s Global Productivity measure (1991:819)) 134
Table 4.6 Negative Prefixes – hapaxes, types and hapaxes/type figures in the BNC 137

Figure 4.2 Histogram showing hapax legomena, dislegomena and trislegomena type-frequency distributions of the negative prefixes – non-, a-, un-, dis- and in- 140

Figure 4.3 Chart showing the negative prefixes over a type frequency range of 1-3 141
Table 4.7 Negative Prefixes – types, frequencies 1-5, 6-10, 11-15, 16-20 and frequency/type ratio 142

Figure 4.4 Histogram showing 1-5, 6-10, 11-15 and 16-20 type-frequency distributions of the negative prefixes 143

Figure 4.5 Chart showing 1-5, 6-10, 11-15 and 16-20 type-frequency distributions of the negative prefixes 143
Table 4.8 Comparing the productivity measures (A, B and C) for the negative prefixes 147
Table 4.9 Comparing the top ten token frequencies in the BNC for un- and a- 148
Table 4.10  Application of the disqualification criteria to filter ‘genuine’ neologisms  
158
Table 4.11  Measure B1 – productivity via a neologism/token ratio  
162
Table 4.12  Measure C1 – productivity via a neologism/type ratio  
162
Table 4.13  Word class breakdown for negatively-prefixed types  
165
Table 4.14  Distribution for negatively-prefixed types  
166

*Figure* 4.6  *Word class percentages for negatively-prefixed types*  
167
Table 4.15  Distribution of negatively-prefixed types by word class  
168
Table 4.16  Word class distribution percentages: negatively-prefixed types vs. whole BNC  
169
Table 4.17  Word class breakdown for negatively-prefixed tokens  
171
Table 4.18  Distribution for negatively-prefixed tokens  
171

*Figure* 4.7  *Word class percentages for negatively-prefixed tokens*  
172
Table 4.19  Word class distribution percentages: negatively-prefixed tokens vs. whole BNC  
173
Table 4.20  Adjectival *A*-prefixed hapaxes in the BNC  
177
Table 4.21  Verbal *Dis*-prefixed hapaxes in the BNC – potential neologisms  
179
Table 4.22  Measure A – word classes – mean type frequency  
180
Table 4.23  Measure A – word class productivity rankings based on a token/type measure for each prefix  
182
Table 4.24  Distribution of the hapaxes for each word class for each prefix  
184
Table 4.25  Measure B1 – word classes – neologism/token ratio  
184
Table 4.26  Measure C1 – word classes – neologism/type ratio  
185
Table 4.27  Domain distribution of hapax sample for each negative prefix and for the BNC as a whole  
187

**Chapter 5**

Table 5.1  The extent to which hapaxes are neologisms and reflect morphological productivity  
206
Table 5.2  Measure C1 - productivity via the neologism/type ratio  
207
ABSTRACT

This thesis investigates morphological productivity, which is among one of the most contentious issues in the field of Linguistics (Aronoff, 1976; Plag, 1999; Bauer, 2001). The advent of electronic corpora sparked fresh interest in productivity, notably the highly innovative measures proposed by Baayen and his co-writers (Baayen, 1989; Baayen and Lieber, 1991; Baayen and Renouf, 1996) which advocate the use of hapax legomena (once-occurring words in a corpus). This research re-investigates these methods to pinpoint the degree to which hapaxes represent new words of the language and, therefore, are valid indicators of productivity. The study also develops a more fine-grained method for measuring productivity, based on the re-analysis of the previously-tried methodologies.

The case study for this research is negative prefixation. A database of all negatively-prefixed words in the British National Corpus is created and relevant variables, such as types, tokens and hapaxes, are examined to determine which combination provides the most pertinent measure. The thesis also comments on the relative productivity of the negative prefixes and therefore considers appropriate historical and semantic factors.

Findings show that though hapaxes do provide access to neologisms, they are not the same phenomenon and cannot, therefore, be seen as a direct reflection of productivity. Significantly, it was discovered that while hapaxes typically include a proportion of neologisms, this proportion varies considerably from affix to affix. The proposed new formulae for measuring productivity - neologism/type and neologism/token - both point towards the prefix non- being the most productive of the five investigated, while in- is the least productive. Findings are also presented about the distribution of the negative prefixes across word class, an area which has never previously been examined in terms of morphological productivity.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and s/he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Head of School of Education (or the Vice-President) and the Dean of the Faculty of Life Sciences, for Faculty of Life Sciences’ candidates.
Acknowledgements

From an academic perspective, I am totally indebted to Steve Jones, who, at the outset, stuck up for me when others were not so keen to. The fact that I got to this stage is completely down to your faith in me and I am very grateful. At the start of your supervisory role, your expertise and knowledge for ‘one so young’ astounded me, and still does. Your guidance and support has been first class, and you instinctively knew when to leave me to ‘get on with it’ and when to ‘give me a nudge’.

Thanks, as ever, to Paul Livesey for his unrivalled expert grammatical knowledge, given so generously over the years. What he doesn’t know about grammar isn’t worth knowing. I am also grateful to Adam Kilgarriff for kindly allowing me access to his Sketch Engine software. The data gathering and analysis was made so much easier by its use. Thanks also go to Heok-Seung Kwon for so generously sending me, from Seoul, South Korea, a bound copy of his doctoral thesis. It has been invaluable. And to Professor Laurence Horn for his incredibly prompt help when I drew a blank.

I would also like to thank Gareth Owens of Heraklion, Crete, both from an academic and a personal perspective, along with Kallia, Michael and David, all of whom in their own way have helped me get to this point.

From a purely personal perspective, thanks go to my mother, Alice Hulse, for her love and support and unquestioning encouragement, and thanks also to my father, George Hulse, who, though he died before I started this current research, gave me my obsession with language and made me the pedant that I am. I also value all the interest shown over the years by close friends, Chris and Mick Taylor, who have been unfailingly supportive. I especially wish to thank Gill Broadbent for keeping me focused, but for reminding me that friends are far more important than any PhD. Despite her own recent setback, her terrific and constant encouragement has been incredible from start to finish, and (since we share a brain) I would like to dedicate half of this thesis to her.

In a perverse sort of way, thanks to Amanda - if you’d not done what you did, I’d never have done what I’ve done.

And finally, σέναν πολύ ξεχώριστο άντρα, ο οποίος ήταν πάντα διαθέσιμος με αγάπη, υποστήριξη και ενθάρρυνση...... ο άνεμος κάτω από τα φτερά μου.
Productivity in Morphological Negation: a corpus-based approach

Introduction

Productivity is the ability of language to coin new complex words from the existing components of the language, typically conforming to the morphological rules of that language. In this way, productivity is the lifeblood of language, allowing it to remain fertile and dynamic. However, the issue of morphological productivity has arguably been affected less by the ‘corpus revolution’ than many other areas of linguistic research. Some would suggest that this is because productivity does not lend itself as readily to corpus-based analysis as other, more context-driven phenomena. What this thesis does, however, is to re-appraise the approaches of those scholars who have exploited corpora to examine productivity, with a view to replicating, refining and enhancing these earlier innovative methodologies.

The principle which underlies almost all corpus-based research into morphological productivity is that those affixes most closely associated with low frequency items are likeliest to be the most productive (Baayen, 1989; Baayen and Lieber, 1991; Baayen, 1992; Van Marle, 1992; Schreuder and Baayen, 1994; Baayen and Renouf, 1996; Plag, 1999). In particular, hapax legomena (once-occurring words in a corpus) have been investigated extensively, since they are thought to include high proportions of neologisms, and neologisms, in turn, are regarded as the best indicators of the productivity of any word-formation process.
This research determines whether the measures applied by Baayen and his colleagues are indeed the most appropriate for the investigation of morphological productivity. The case study presented is that of negative prefixation, a means of creating new words not previously afforded any great attention in the literature. Specifically, this research aims to:

1. to replicate established methods for assessing productivity using more substantial corpora;
2. to assess the degree to which hapax legomena are neologisms and therefore suitable indicators of productivity;
3. to compare the different formulae used in the calculation of morphological productivity, and consider which is most fit for purpose;
4. to shed light on the relative productivity of negating prefixes in English and to determine which are currently most productive;
5. to develop a more fine-grained way of measuring morphological productivity based on the findings above.

The thesis begins by placing the research in context, and describes the notion of productivity, examines its scope and importance in language, and discusses the complexities of the phenomenon. Chapter One continues by contextualising negative prefixes as the case study in this research, and discusses a number of views which place negation in a pivotal role in the human experience, and justifies the choice of negative prefixes over other affixes. The chapter then introduces the corpus approach which allows authentic language data to be examined and which, to date, remains the most comprehensive means of accessing huge quantities of naturally-occurring language.

The literature which relates to language productivity, negative prefixation, and the use of electronic corpora is examined in Chapter Two. In terms of productivity measures, due to their development at different periods of time, the
chapter deals with previous studies in a chronological fashion, since this particular research does not lend itself to the normal literature review format. The chapter discusses pre-corpus approaches, such as intuitive and dictionary-based accounts, and then moves to more recent corpus-based methods. The chapter recognises the contribution to corpus-based methods of Baayen and the various other linguists with whom he has worked, and broadly agrees with the hapax-based approach which Baayen promotes. It concludes, however, that more refined measures are required to assess the usefulness of hapax legomena in a measure of productivity.

Chapter Three describes the data collection procedures, building the database on which this research is founded. It justifies the use of the British National Corpus (BNC) as the basis for the creation of the database, and also the choice of the five negative prefixes to be used. The chapter also discusses the manual element involved in editing the database and the subsequent type and token tables created for use in the following chapter. Chapter Three also introduces some of the historical and semantic factors which affect the use of the negative prefixes.

Chapter Four involves the collection of the statistical evidence and a close examination of the BNC’s hapax legomena, and subsequently goes on to discuss measures based on type and token frequencies, along with measures based on the presence of hapax legomena, and investigates the degree to which hapaxes reflect productivity. The chapter also analyses and discusses the influence that word class has on the productivity of the negative prefixes and, in addition, it moves towards proposing new measurements of productivity.

Finally, Chapter Five presents an evaluation of the results of the study and sets out the ways in which this research has taken the notion of productivity forward. The
Limitations of the study are also presented, along with recommendations for future research.
Chapter 1

Research Context

The aim of this research is to examine the phenomenon of language productivity, and specifically to investigate how productivity is measured. To this end, the study replicates, and improves upon, earlier research methods into measuring productivity, and uses the British National Corpus (BNC) as the basis for its data gathering. The overarching aim is to produce a finer-grained measurement system to measure morphological productivity. The specific area of investigation is the notion of language productivity, and it utilises a case study based on negative prefixation, which is to be analysed in terms of a corpus-based approach. This initial chapter delineates the scope of the thesis, and discusses the choice of topic and the rationale for this choice.

Bauer examines productivity from the perspective of the processes involved and proposes that “any process [...] is said to be productive if it can be used synchronically in the production of new forms, and non-productive if it cannot be used synchronically in this way.” (1983:18). Plag considers that it is the affixes which are the productive elements and proposes that the “property of an affix to be used to coin new complex words is referred to as the productivity of that affix. [...] Even among affixes that can in principle be used to coin new words, there seem to be some that are more productive than others.” (2003:44). This apparent distinction in the definitions relating to productivity is, however, “more apparent than real. Productivity can, at all times, be considered to be a feature of individual morphological processes.” (Bauer, 2001:32). Indeed, Crystal broadly defines
‘productivity’ as a linguistic term which refers to the creative capacity of language, and goes on to state that:

The term is also used in a more restricted sense with reference to the use made by a language of a specific feature or pattern. A pattern is productive if it is repeatedly used in language to produce further instances of the same type. [...] Non-productive (or unproductive) patterns lack any such potential [...] Semi-productive forms are those where there is a limited or occasional creativity, as when a prefix such as un- is sometimes, but not universally, applied to words to form their opposites, e.g. happy → unhappy, but not sad → *unsad.

(Crystal, 1997a:310)

This thesis, therefore, examines productivity in this more restricted sense proposed by Crystal (1997a), and investigates how productivity is measured. It uses negative prefixation as a case study and applies a corpus-based approach to the investigation.

1.1 Introduction

So far, we have established that productivity is the capability of language to coin new complex words from existing components, which generally conform to the morphological rules of that language. As such, productivity allows a language to remain alive, fertile and dynamic. Productivity is, therefore, a fundamental requisite of language; an obligatory feature.

The concept of productivity in language is a long accepted one. As Schultink suggests “the whole notion of grammar implicit in the work of the Sanskrit grammarians assumes the idea of productivity.” (1992:188, cited in Bauer, 2001:11). It is, therefore a long-established phenomenon, yet the concept still remains one of the most controversial and disputed issues in linguistics. Mayerthaler, for example,
describes productivity as “among the most opaque concepts of linguistics.” (1988:92), and Weinreich, who writes the foreword to Zimmer’s 1964 monograph, proposes that “…linguistics has made do with a conception of productivity much flabbier than a science of language is entitled to.” (3).

Superficially at least, the ‘average’ English speaker would appear to know nothing of the science of language, of productivity, of the structure of words, of the creation of new complex words, or adhering to the morphological rules of the language. But in fact we do and we do this all the time. We produce new words to fulfil an immediate need, and also possess an intrinsic ability to know how to combine the elements which make up words. We also come across (often newly-created) words in the media and the press, almost on a daily basis, that we have never encountered before, yet are able to understand the meaning of such words as we recognise the meaning of the component parts. Similarly, if we encounter the words symmetrical (adjective), honesty (noun), legally (adverb), smoker (noun), and cover (verb), as English speakers, we would have little difficulty in producing their negative forms, even though we would most likely choose different negating elements for each example (asymmetrical, dishonesty, illegally, non-smoker, and uncover). In other words, we know from experience which negating element will combine with each base word. This thesis explores these intuitions more closely, to specifically discover how speakers decide which negating prefix to use when formulating a new word. The current chapter then follows a three-part structure, and specifically focuses on the three key elements of the research. Firstly, productivity is examined in this chapter as the thesis is ultimately about productivity. Secondly, negation is examined as morphological negation constitutes the case study of this
research, and thirdly, corpora are discussed, since the methodologies utilised in this study are corpus-based.

1.2 Productivity

The initial section of the thesis has outlined the differences in the definitions which relate to the notion of productivity. These definitions of productivity and how productivity is demonstrated can now be explored in more detail.

Productivity is a vital feature of any language, yet defining, describing and quantifying productivity remains ambiguous, and persists as one of the most contentious issues in the study of Linguistics. The diversity of opinion with regard to productivity is hardly surprising, since no common fundamental baseline has been established as a starting point for its evaluation. Productivity appears to mean different things to different people, Anderson, for example, like Bauer (2001), views processes as the productive (or non-productive) elements (1982:585), whereas Aronoff argues it is the rules of the language which are productive or not (1976:36). The main disagreement among linguists appears to relate to the degree of productivity of the feature involved. Bauer argues that one approach (amongst several he puts forward) is to describe productivity as the ‘potential’ of a rule-based process for repetitive coining (2001:99). Whilst it can be acknowledged that a word formation process may be latent and have unrealised ability to coin new words, a definition of productivity in terms of ‘potential’ nevertheless appears to deny any recourse to quantitative measures.

Variety of opinion similarly occurs with regard to whether productivity in itself should be considered a binary or a gradable phenomenon. For some, such as Booij (1977) and Zwaneburg (1983), productivity is a binary system; morphological
processes are either productive or not, with no middle ground. For others, such as Matthews (1991), Bauer (1992b) and Plag (1999), the phenomenon forms a continuum, and “completely unproductive or fully-productive processes only mark the end-points of a scale.” (Plag, 1999:11-12).

The terminology which relates to productivity is also the subject of differences in opinion. What some refer to as ‘productivity’, Chomsky, for instance, calls ‘creativity’ (1965:6), and his view of the phenomenon includes syntactic and inflectional processes, in addition to that of derivation. This is ‘productivity’, as Di Sciullo and Williams claim, “in the most basic sense of the word” (1987:8). For others, such as Lyons, there is a clear distinction between ‘productivity’ and ‘creativity’. Lyons proposes that ‘productivity’ is the property of a language system that allows speakers to “construct and understand” a huge number of words which they have never encountered before (1977:76). On the other hand, Lyons argues that ‘creativity’ relates to “originality of style; and it is for this kind of novelty or originality that the term ‘creativity’ is most appropriate (though it is not uncommonly used by linguists for what we are calling productivity)” (1977:77). Haspelmath submits that several linguists treat the distinction between ‘productivity’ and ‘creativity’ as the difference between unconscious, unintentional coining and intentional formations which use otherwise unproductive processes (2002:100).

There are, in addition to terminological discrepancies, a number of different perspectives from which productivity is viewed and has been defined in the literature. Haspelmath, for instance, puts forward the view that the productivity of a process can be determined by the lack of restriction on that process, and presents the idea that “all morphological rules are equally productive, but they are not equally restricted. Some are quite unrestricted […], whereas others are heavily restricted”
For example, the suffix -ness will attach to a vast array of adjectives (as well as to participles, adjectival phrases, and occasionally adverbs, nouns, pronouns, and verbs), almost without restriction, to form abstract nouns, such as, selfishness, bitterness, helplessness, greatness, etc. (examples taken from the OED Online, 2008). On the other hand, the suffix -en which forms verbs from adjectives, such as blacken, darken, harden, widen, etc., only attach to monosyllabic adjectives which end in a plosive or fricative sound (Haspelmath, 2002:111). Haspelmath, however, rejects this notion on the basis that “[t]here are simply too many rules that are not obviously restricted heavily and yet their productivity is limited”, and cites the diminutive suffix -let as an example of an affix with few restrictions that is rarely used in the creation of new words (2002:111).

Another view of productivity is proposed by Aronoff who argued that the productivity of a process should not merely be viewed as a historical phenomenon (1980:81). Instead, he proposes that a description of productivity should also take into account the speakers’ synchronic competence of the language, since speakers have been “found to be sensitive to the productivity of the word formation rules” (1980:81). Haspelmath (2002) reports Aronoff’s findings, which suggest that when speakers are presented with ‘potential’ new words which incorporate ‘productive’ and ‘unproductive’ affixes, “their acceptability judgements strongly correlate with the productivity of the rules” (111). Aronoff believed that, up to this point, the productivity of a process was only measured in terms of words which had already been ‘produced’ by a process, but did not convey any indication of the ‘potential’ of the process (1980:72). He, thus, proposed a judgement measure of productivity based on “synchronic descriptions of linguistic competence” (1980:81).
A further perspective offered is that of semantic and phonological transparency (Haspelmath, 2002: 111). In this view, the meaning and pronunciation of a neologism must be transparent, since otherwise it would not be understood, and this must, therefore, indicate at least a correlation between productivity and transparency. Haspelmath presents a further proposal that a correlation most likely exists between the productivity of a process and the number of actual words produced by that process. However, ultimately, his position is that, while productivity is often seen as a phenomenon of language use (in other words, performance, it is in fact language competence which is the relevant variable in measuring productivity (2002: 112). Haspelmath proposes that, unlikely as it sounds, speakers obviously do observe and (mentally) record neologisms, and states that “our linguistic knowledge comprises not only what one can and what one cannot say, but also what one is likely to say”, certainly in terms of neologisms (2002:112). This raises issues, however, as to how and when, and by whom, the existence of a new word is documented, and equally important, how the number of words are measured. All of these issues are discussed in subsequent chapters.

1.2.1 The scope of productivity

As with defining and categorising productivity, opinions vary as to the scope of productivity. In particular, disagreement arises as to whether processes such as clippings (porn from pornography), blends (motel from motor and hotel), acronyms (VAT [pronounced /væt/] from Value Added Tax), back-formation (self-destruct from self-destruction), compounding (armchair from arm + chair) and, more critically, inflection (talked from talk) should be treated as productive processes. For many linguists, such as Schultink (1961), Aronoff (1976), Booij (1977), Bauer (1983;
1988), Van Marle (1985), and Baayen and Lieber (1991), clippings, blends, acronyms, back-formation and compounding are excluded from productivity measures on the basis that they are not typically determined by a rule-governed system (Leech, 1981:221; Van Marle, 1985:45; Carstairs-McCarthy, 2002:94). For them, they are not regular word formation processes, which incorporate simplex bases (bases composed of a single part or structure) and affixes (bound morphological components other than the base word, which cannot occur in isolation) in combinations which are, in theory, infinite. Whilst compounding in English (e.g. *nosebleed*) can be classed as a word formation process, again, it is not specifically associated with a rule-led system and does not typically or necessarily incorporate affixes.

Inflection, on the other hand, is determined by a rule-governed system, and does incorporate simplex bases and affixes. However, inflection does not create an entirely new word; it involves the modification\(^1\) of a word to mark some form of grammatical agreement. Inflection is a closed system, which is semantically regular, whereby the modified word always retains its word class - it keeps its essence - and the meaning of the modified word remains the same as the stem from which it was formed. In the case of derivational word formation, a rule-governed process is involved; simplex bases and affixes are the components in an open, theoretically infinite, system; but the end product is a different word to that of the base word, which may or may not have involved a change of word class – for example, *happy >> unhappy* (adjective >> adjective) and *hope >> hopeful* (noun >> adjective). The position taken by this research is that productivity relates to rule-governed, derivational, word formation processes, which involve bases and affixes in an open

---

system, which may or may not change the word class of the base, but always produces an entirely different lexeme, as opposed to a different form of the same lexeme.

### 1.2.2 The importance of productivity in language

As referred to in Section 1.1, for the most part, the ‘average’ speaker of a language is unaware of the significance of productivity. Nevertheless, they are perfectly able to use the components of the language productively to generate new words. In turn, the hearer can generally understand the combination of these components when they come across them. For example, whilst we may have never encountered the word unsearchability, we would almost certainly be able to assume its meaning from the familiar components with which it is constructed (*un*-search-able-ity), since we will have encountered the individual components in other existing words. We are, in fact, immersed in the productive use of these components in newspapers, on television, in advertising, in texts of all types, and particularly in terms of new technology and the use of the Internet. “Productivity […] therefore cannot be overlooked in a correct description of a linguistic system, and the linguist who neglects this particular factor will be counting ‘dead souls’ as live people.” (Marchand, 1969:5).

### 1.2.3 Areas of relevance to productivity

As a fundamental phenomenon of human language, productivity is pertinent to a variety of linguistic domains. For example, as part of the National Curriculum in English (www.nc.uk.net), children learn about productivity in terms of their study of affixation and compounding; for instance, about the morphological components

---

2 The positive term searchability, though first recorded in the OED in 1946, has been more recently used “with reference to electronic data” (OED, 2008).
which can be combined to create complex words. Specifically, under the National Curriculum, children learn “the meaning, use and spelling of common prefixes and suffixes” (http://curriculum.qcda.gov.uk/uploads/English [p28]). They are provided with examples of the components, shown how these combine with base words, and are provided with the ‘acceptable’ resulting words. Children are, however, given little or no explanation as to why certain combinations are considered ‘acceptable’. For example, unpleasant, dishonest, or unhappy are ‘acceptable’, whereas other combinations, such as *displeasant, *unhonest or *dishappy, would be ‘unacceptable’ formations (http://www.icteachers.co.uk/children/sats/prefixes.htm). In short, pupils are told what goes with what, as opposed to why. A similar lack of explanation also occurs in terms of learning English as a foreign language.\(^3\)

One significant example of innovation in contemporary language is in the area of technology, which could be considered a morphologically productive genre. As new technological cultures continue to develop, new words will most certainly be needed to describe new entities and actions. Today, the age of the computer and, more specifically, the phenomenal increase in the development and use of the Internet, has meant that the creation of new terminology has been inevitable. In some cases, this has been achieved by the utilisation of existing words in new metaphorical and analogical senses; for example, mouse, file, folder, icon, menu, etc. But equally important, new technology has led to the production of entirely new terminology, which in turn, has led to the creation of new morphologically-complex words. For instance, the verb unsubscribe has been created from the existing verb subscribe. Contemporary language has taken existing words which, until recently, have had a primarily literal sense and put them to metaphoric use. Due to the nature of this

\(^3\) See also the following websites http://www.a4esl.org/q/h/vm/negpref.html and http://www.bbc.co.uk/world_service/learningenglish for National Curriculum and foreign language learning details.
metaphoric use, these words have, in turn, become available for further derivation. The verb install can be used to illustrate this point. Install has a literal sense of ‘physically placing something in position for use or service’, generally as a permanent fixture. Install has subsequently been used metaphorically in the sense of ‘to add a computer application’. Since this is a virtual application, it is not necessarily a permanent feature, and is, therefore, removable. As a result, we now have the negated term uninstall, meaning ‘to remove a computer application’.

Where there are no currently-available terms to describe an entity or a process, we are perfectly capable of producing our own new word. In some situations, this is merely due to the fact that no terminology has previously existed to precisely describe our meaning. Alternatively, productivity may occur purely to capture a more explicit sense than any existing term may allow. For instance, see below real-life and media examples compiled over the period that this research was undertaken:

“That was a non-joke.”
(Chris Taylor, 11/03/05 – discussing a joke that was not amusing).

“Do you mean a non-stick one or a non-non-stick one?”
(Alice Hulse, 30/04/06 – discussing the purchase of a new wok).

“The session on Thursday has been un-cancelled.”
(Mary Faith Autumn, 28/11/07 – discussing a previously-cancelled sign-language interpreting session).

“Well, you’ll just have to un-promise it.”
(The character Janice Battersby 03/10/08 – discussing the promise of money in the ITV series Coronation Street).

“I said I’d bring you or they’re probably going to un-invite me.”
(The character Lucy Beale, 09/01/09 – discussing a party invitation in the BBC series Eastenders).
“I’m going to be drinking lots of non-alcoholic lager and getting seriously undrunk.”
(Jane Brunning, 07/08/03 – celebrating a success, but being allergic to alcohol).

Though these coinages may never become established words of the language, nonetheless, new words were produced to fill a lexical gap. In the case of the example ‘undrunk’, for instance, whilst there is an existing word which describes the process of becoming intoxicated, there is no equivalent term to describe the process of drinking without specifically becoming intoxicated, yet the meaning conveyed by the production of the new word was readily understood by the audience in the given context.

The study of productivity is of importance to the field of lexicography, since a dictionary compiler must have up-to-date information which relates to what is currently acceptable and what is not. The lexicographer requires in-depth knowledge of the prevailing trends in language. In addition to listing headwords in a dictionary, along with their potential variety of polysemous senses, and possible word class variations, the semantic profiles of affixes and combining forms also need to reflect current usage and acceptability. What was unacceptable only a few years previously may now be perfectly acceptable, and what was perfectly common in the past, may have now become infrequent or even obsolete.

The negative prefix in-, for example, was formerly only attached to words derived from Latin, whereas un-, as an indigenous prefix, could be freely attached to words with a variety of etymologies to create the negative form. In more recent years, the prefix un- has additionally spread to words which were formerly only prefixed with in-, particularly in cases where the prefix in- has developed a more specific sense, as opposed to a purely negative one. For instance, the examples
inhuman and unhuman, though both represent a type of opposition to the word human, the two carry different senses. Unhuman can mean “not limited by human qualities or conditions; superhuman [...] not pertaining to mankind” (OED), whereas inhuman has developed the narrower sense of ‘lacking humane feelings, cruel and brutal’. This is not, however, always the case, and words prefixed with in- do not always have a more specific sense. Examples such as cautious, certain, communicative and distinguishable can take either prefix without any semantic distinction, though nowadays un- would typically be the preferred choice of negative prefix. “(T)he modern tendency is to restrict in- to words obviously answering to Latin types, and to prefer un- in other cases” (OED). Similarly, the negating a- prefix came into English from Greek and was, until comparatively recently⁴, only attached to words of Greek origin, for example asymmetrical. The prefix is, however, now also used to negate words of Latin origin, for example, amoral, asexual, and asocial. Indeed, Baldi et al (1985) propose that the etymology of the base word plays a significant part in the choice of one negative prefix over another, and that in fact native speakers do appear to “have access to and make use of etymological information about roots and prefixes before assigning prefixes to roots” (34). This issue is returned to in Chapter 3, Section 3.4 with regard to how historical factors affect the negative prefixes.

1.3 Negation

The previous section examined the importance of language productivity with specific reference to the need for new terminology as a result of developing technological

---

⁴ According to the OED Online, the first recorded negation produced in English using the negating a-prefix was during the Renaissance in 1646, when the ‘negative’ noun adiabolist was formed from the positive term diabolism. It was not, however, until 1835 that the prefix was added to a word of Latin origin, normal to produce abormal, though this was subsequently altered to the now more familiar abnormal (OED Online, 2001).
cultures, along with the need for innovative, productive use of language components to describe new entities or processes, or indeed, to accommodate new, perhaps more explicit, senses for existing words. Because this thesis uses negative prefixation as its case study, the current section justifies why morphological negation is an important and profitable topic in a discussion of productivity. In describing negation, Horn, for example, proposes that it is our ability to negate that is one of the distinguishing features that separates human language and animal communication, and that this distinction:

....can be argued to result directly from the essential use humans make of negation and opposition. If we are by definition the animals that talk, we are ipso facto the animals that deny, for as Spinoza and Hegel argue, any linguistic determination directly or indirectly involves a negation.

(2001:xiii)

Likewise, Plato claims that, “….objects never purely possess one property rather than its opposite.” Even Socrates, of whom Plato was a devoted follower, is said by Plato to be both just, but in some respects, also unjust (Plato, via Penguin Dictionary of Philosophy, 1996:426). Thus, in terms of logical reasoning, our capacity to negate is considered a vital characteristic of human behaviour and experience; such that viewing the world in terms of opposite conditions and contrast appear to be a basic universal human tendency. As Horn proposes, “(n)egation may or may not be the most basic, most debated, or most ancient of the logical connectives, but it is without doubt the most maligned.”, and on this basis warrants further investigation (2001:45). Whether we accept the view that the ability to negate is a fundamental human trait, or whether we believe that negation is simply an inherent characteristic of language, is not important. Whichever position is taken, our ability to negate and
to produce opposition in language “is one of the most important principles governing
the structure of languages” (Lyons, 1977:270).

1.3.1 What is negation?

The capacity to negate is the capacity to refuse, to contradict, to lie, to speak ironically, to distinguish truth from falsity - in short, the capacity to be human.

(Horn, personal communication 10/05/10, taken from his blurb for Horn, 2001)

Despite its seemingly pivotal position in human behaviour and experience, the phenomenon of negation can be defined in a number of different ways in English. It is an abstract phenomenon whereby some form of contradiction or opposition is expressed, though this is achieved either grammatically or semantically, which is realised by a specific grammatical construction or by the application of a specific process.

Often in English, negation is achieved sententially by the addition of the negative particle not (or its contracted form n’t) to the primary verb (e.g. ‘He is not tall’), or to the first auxiliary verb (e.g. ‘She wouldn’t have been going anyway’), and the scope of the negation then conventionally continues from the negative particle to include the remainder of the clause or sentence. Morphologically, however, negation in English is realised by the attachment of certain negative prefixes, such as dis-, non-, un-, etc. to the base word, though the addition of a negative prefix negates only the base word and does not grammatically affect the remainder of the sentence.
Tottie (1980), for example, describes this distinction as ‘non-affixal’ (sentential negation) and ‘affixal’ (morphological negation) respectively.\(^5\)

### 1.3.2 Morphological negation as a case study

The world *can* be described without the use of the word ‘not’.\(^6\)

(Bertrand Russell, 1948:520, cited in Horn, 2001:50)

Given that negation appears to be such an important feature of human behaviour and experience, it seems surprising that one major aspect of negation is given far less attention than another. Research into grammatical (non-affixal; sentential) negation is far more widespread than research which relates to semantic (affixal; morphological) negation, and this section, therefore, justifies the choice of morphological negation as the case study for this thesis. In relation to language productivity, morphological factors are more pertinent than the grammatical features of a language. Describing and determining grammatical rules, though significant in providing a grammar of a given language, are not so relevant in terms of how those rules are used, whereas evidence of ‘actual’ usage of morphological constructions must surely remain of prime importance in any study of productivity.

Morphological and sentential negation may appear as synonymous and semantically equivalent. Superficially at least, there seems little, if any, semantic distinction between the different types of negation. Compare: ‘*A bachelor is a man who is not married*’ with ‘*A bachelor is man who is unmarried*’, whereby *(not)*

---

\(^5\) It should be noted that Tottie (1980) further sub-divides ‘non-affixal’ negation into ‘analytic non-affixal’ *(not)* and ‘synthetic non-affixal’ *(‘never’, ‘no’, ‘none’, ‘nothing’, ‘nobody’, ‘nowhere’, ‘nor’ and ‘neither’)*, and within the ‘affixal’ category, she also includes the negative suffix *(–less)*. For the purposes of this study, the ‘non-affixal’ category is limited to the ‘analytic’ variety, and the ‘affixal’ category limits itself to the prefixal variety.

\(^6\) My italicising.
married and unmarried make up a pair of non-gradable adjectives, and there is little, if any, difference between the two semantically. Tottie (1980), however, claims that any “(l)ack of synonymy occurs chiefly with gradable adjectives” (103). She goes on to propose that it is a range of semantic constraints which ultimately influence our selection of morphological or sentential negation (1980:103)\(^7\). Despite a potential lack of synonymy, the distinction between morphological and sententential negation appears relatively uncomplicated; however, it is the resulting ‘degree’ or ‘quality’ of negation from each type which is less explicit. For some, such as Ayer (1952, via Horn, 1989) and Jackendoff (1969), morphological negation is not ‘true’ negation, but instead makes use of “privative statements” which are in fact “affirmative in nature”, as opposed to negative, for example, Michael is unhappy (Horn, 1989:33). Others, such as Jespersen (1917:144), believe that morphological negation is a ‘purer’, ‘simpler’ form of negation than the sentential variety. Jespersen (1917:144), Zimmer (1964:23) and Horn (1989:33), however, do agree that morphological negation typically produces contrary opposition (gradable), as opposed to a true contradictory opposition (binary; non-gradable). Jespersen (1917:43) and Zimmer (1964:23) also concur with Lakoff and Johnson (1981:129-130) that the morphological variety produces a stronger and less ambiguous form of negation.

Sentential negation has been well covered in the literature, but, until recently, Zimmer’s 1964 monograph provided perhaps the only significant work which deals specifically with morphological negation. Others, such as Marchand (1969), Funk (1971), Kwon (1997), and Hamawand (2009) have dealt with lexical negation more generally, purely semantically, and diachronically and, since two of these works were produced around forty years ago, before the widespread use of electronic

\(^7\) See also the views of Jespersen (1917:42) and Horn (1989:331).
corpora, the findings were inevitably based on ‘possible’ rather than ‘actual’ usage. A study of morphological negation in terms of ‘actual’ usage, therefore, appears to be long overdue.

Specifically, therefore, this research focuses on morphological productivity from the perspective of negative prefixation that can be achieved by the incorporation of a number of different negative prefixes, which in some cases stand in valid competition with each other. So, for example, from the adjective *moral*, the negative forms can be *amoral*, *immoral*, *non-moral*, and *unmoral*. Likewise, from the adjective *typical*, the negative forms *atypical*, *non-typical*, and *untypical* can be created. It is this multiplicity of means to negate and the potential subtle semantic variance which make negative prefixation such a suitable (and fascinating) case study and, it is the belief of Baldi et al, for instance, that, in terms of morphological productivity, “(n)egative prefixing is a fertile testing ground” (1985:33).

1.4 **Corpus Linguistics**

By necessity, studies of productivity prior to the 1980s were based on intuition, or on methods involving judgement or preference testing of subjects. Whilst intuitive notions have often been subsequently verified, and judgement and preference testing have certainly revealed the overall views of the specific subject base tested, up to the 1980s, there were no means of testing a wider, more representative, population. Since the 1980s, and with the event of electronic corpora, the means of accessing vast amounts of data became available. This research, therefore, utilises an electronic corpus approach in order to measure the use and productivity of the negative prefixes.
“In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description” (Kennedy, 1998:1). Thus, a corpus is a tool which can be employed to help illustrate patterns and trends in language. It could be argued that a corpus approach does not in itself make specific and definitive statements or judgements about language, but it does allow observations to be made which, in turn, can support or refute theories of language use. The utilisation of corpora is an empirical approach, which allows analysis of authentic data drawn from attested examples of language. In other words, a corpus makes use of ‘real’ language as its foundation, which also facilitates both qualitative and quantitative analysis. Electronic corpora provide the means to examine huge amounts of data, which previously it would have been inconceivable to handle, store or analyse (Biber et al, 1998:4). The immense quantitative capacity of an electronic corpus allows more reliable generalisations to be made about language, which in turn facilitates more robust explanations in terms of qualitative interpretation. By their nature, electronic corpora reduce the errors and restrictions unavoidable with manual searches and data analyses, and provide statistically reliable and exhaustive information which can be readily and accurately replicated. Characteristically, “The corpus provides contexts for the study of meaning in use….” (Kennedy, 1998:9).

1.4.1 Attitudes to corpus-based approaches

The use of corpora has become more commonplace and more accepted over recent years, and their reliability as tools to aid language description and analysis has greatly improved over time. Corpora should not, however, be viewed as the ultimate resource in language analysis, since qualitative analysis is equally as relevant as
quantitative measures. Indeed, “(s)ome of the most revealing insights on language and language use have come from a blend of manual and computer analysis.” (Kennedy, 1998:2). Corpus-based measures are tools to be utilised and, as such, need to be appropriate for the job in hand. In other words, a corpus needs to be large enough and carefully constructed, yet varied enough, to permit generalisations to be made about the language and, in turn, about productivity. “Corpus linguistics is not an end in itself but is one source of evidence for improving descriptions of the structure and use of languages” (Kennedy, 1998:1). Thus, a corpus-based approach, accompanied by a degree of manual input and qualitative analysis, appears to be both an ideal starting point, as well as being a synergic combination for an investigation into morphological productivity, in order “to discover facts [...] which have never been noticed or commented on previously.” (Kennedy, 1998:3).

In the early days of electronic corpora, criticisms were levelled at corpus-based studies, not least by Chomsky, who argued that a corpus could not ever be representative of language. He claimed that a corpus would be inevitably biased, distorted, and thus unrepresentative, as well as being out-of-date from the moment it is compiled.

Any natural corpus will be skewed. Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.

(Chomsky, University of Texas, 1962:159, cited by Leech, 1991:8)

Presumably, however, this would also be the case with any body of evidence, since all sources must contain some degree of subjectivity given the need for some degree of judgement and intuition with regard to input. As a result, all evidence must be biased and dated to some greater or lesser extent. The fact remains, however, that, to
date, the electronic corpus is the most comprehensive means to access vast quantities of naturally-occurring language.

(W)e can see, in historical retrospect, how the availability of vastly increasing computer corpus resources has enabled syntactic and lexical phenomena of a language to be open to empirical investigation on a scale previously unimagined.

(Leech, 1991:13)

It might be argued that corpora can, at best, only reveal tendencies and trends in language, or very general paradigms, but that they reveal little of a system’s rules, or alternatively that statistical evidence from corpora is of little relevance in enriching our knowledge of language. In describing the dichotomy between the corpus linguist and the “armchair” linguist, Fillmore wittily suggests that:

These two don’t speak to each other very often, but when they do, the corpus linguist says to the armchair linguist, “Why should I think that what you tell me is true?”, and the armchair linguist says to the corpus linguist, “Why should I think that what you tell me is interesting?”

(1992:35)

Yet corpus evidence is just that; it is evidence, it reveals what is happening in language; and it utilises ‘real’ attested language as its source material. Language changes; and it is not beyond the realms of possibility that what is considered inadmissible in the language today may, for various reasons, be considered entirely acceptable in the future. For example, originally coming from Latin, the adjective moral would formerly be negated by the addition of the prefix in-, which gives the negated form as immoral. Subsequently, due to changes in the acceptability of combining Latin bases with non-Latin prefixes, and potentially poly-semantic uses, amoral, non-moral and unmoral are now all permissible and carry discrete, if not always conspicuous, meanings. Linguistically, there seems little to be gained from the investigation of the status quo in terms of acceptability, when we continually
develop new senses for formerly mono-semantic words, and also adapt and revise what is permissible, as indeed we must in order for language to remain dynamic. What we *ought* do linguistically (i.e. prescriptively) and what we actually do in reality, rarely concur, and one advantage of corpora lies in their ability to recognise this distinction between the theory and practice of linguistics.

Thus, identifying trends, tendencies, patterns and paradigms are all the goal of the corpus linguist. They seek to make generalisations about language and language change. As Kennedy argues:

> …generalizations are an essential part of science and we have no difficulty accepting generalizations about the human body […] even when we know that every person’s body is different […] In phonology, we have little difficulty accepting generalizations about the sound systems of a language even when every speaker of the language sounds different. The great dictionaries and grammars of English are all generalizations about the language in this sense.

(1998:62)

1.4.2 The rationale for using a corpus

Without exception, every non-linguist who has enquired about the topic of this research, and has been told that it concerns negative prefixes, has responded with something which approaches “Oh, you mean like *un-*?” Certainly, *un-* appears to be the prefix which comes to mind most readily for most non-linguists. Note also that, by consulting a standard dictionary, this intuition appears to be confirmed. The *Concise Oxford Dictionary* (1996), for instance, lists a total of 86 words negatively-prefixed with *a-*; 47 words negatively-prefixed with *dis-*; 336 words with *il-, im-, in-* or *ir-*; 113 words with the prefix *non-*; but a total of 1,262 words negatively-prefixed with *un-* (Hulse, 2002:97). This strongly indicates that the prefix *un-* is by far the most common, and potentially, therefore, the most productive of the negative prefixes. As a case in point, when comparing *un-* to the negating *a-* prefix, Bauer
states that the \textit{a-} prefix is “….losing ground to the much more productive \textit{un-}….\)” (1983:218), and goes on to propose that “….the replacement of \textit{atypical} by \textit{untypical} is evidence of the greater productivity of \textit{un-}….\)” (personal communication, 17/2/02). However, in the British National Corpus, a raw frequency of 100 is recorded for \textit{untypical}, whereas a frequency of 144 is noted for \textit{atypical}. This suggests that perhaps our intuition is not as reliable as we might wish to believe. Naturally, this raw data is purely quantitative, but the use of corpora also allows us to examine this data from a more qualitative perspective by means of the availability of contextual information.

Negation - more precisely morphological negation - is inextricably linked with antonymy. In the case of the adjective \textit{natural}, for instance, intuition would suggest that \textit{unnatural} would be elicited as both its opposite and its negated form. As Jones states, “If pressed to give an ‘opposite’ [for \textit{natural}], most speakers would probably opt for \textit{unnatural}.” (2002:160); that is, the morphologically-related negative version of the base word. Yet in his corpus-based investigation, Jones demonstrated that, in a 280-million-word corpus\textsuperscript{8}, \textit{unnatural} did not appear once in textual opposition with \textit{natural} (2002:163), though words like \textit{artificial} and \textit{man-made} appeared on several occasions. The implication here is that the morphologically-related negative version of a word is not always its most-used antonym.\textsuperscript{9} Since learners’ dictionaries, dictionaries of synonyms and antonyms, and thesauruses variously cite both lexical and morphological options to signal opposition, it would seem that only by means of corpora investigation can these alternatives be distinguished in terms of pragmatic and attested usage. There would

\textsuperscript{8} \textit{The Independent} newspaper corpus (1\textsuperscript{st} October, 1988 to 31\textsuperscript{st} December, 1996).

\textsuperscript{9} It should be acknowledged that Jones (2002) utilises the newspaper corpus to “reflect a natural, modern, non-fictional use of written language” (p.26).
appear to be no other means to examine speakers’ choices, at least not on such a large scale as is viable with corpus analysis.

Prior to the now more widespread and accepted use of electronic corpora, much of the pronouncement made about language productivity was, by necessity, based on intuition and speculation. That is not to say that intuition and speculation play no part in corpus linguistics. Some element of the subjective always exists in any theory, and intuition still has relevance in any hypothesis and any analysis. As Jones states, “Corpus linguistics is not intuition-free linguistics.” (2002:44). Any theory is based upon a hypothesis, and ultimately that hypothesis is founded on an intuitive notion. Through investigation and analysis, a possible interpretation can be provided for a phenomenon, and a corpus should be seen as a tool, a methodological resource, which involves the examination of language in use, in order to detect trends and recurring patterns. In turn, this provides meaningful statistics and facilitates credible explanations. Perhaps the ultimate benefit of the use of a corpus approach is its transparency and thus its ability to be replicated, which, if found wanting, can be questioned and re-evaluated. Although it initially appears to be a criticism of corpora, consider the following from Fillmore:

I don’t think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. [But] every corpus that I’ve had a chance to examine, however small, has taught me facts that I couldn’t imagine finding out about in any other way.

(1992:35)

The use of corpora is perhaps the only means to empirically identify patterns and trends in language, and a study of linguistic phenomena can only benefit from such empirical evidence. Certainly, corpora reveal what is happening in the language and allow changes to be identified. As larger corpora become available,
quantitative generalisations become more reliable and more robust. Although this study is not the first to examine morphological negation from a corpus perspective, it is, however, the first to discuss the issue in terms of morphological productivity, utilising a quantitative analysis, in addition to a qualitative element, which the use of corpora facilitates.

1.5 Research context summary

This initial chapter has outlined the areas to be covered in this thesis with regard to the choice of topic and the rationale for this choice. The principal field to be investigated is that of language productivity, and involves a case study based on negative prefixation, which is to be analysed in terms of a corpus-based approach.

This opening chapter has examined the nature and scope of productivity, and has discussed why productivity is an important issue in language. It has been suggested that productivity is a particularly relevant theme when examining the language of contemporary culture, which is, in turn influenced by fast-developing technology, whereby new words are needed and, therefore, created to fill lexical gaps. Likewise, lexicology and etymology have also been considered with regard to current productivity and current trends in the language.

As this thesis deals with negation as its case study, this chapter has defined negation, and has distinguished between morphological (prefixal) and sentential (non-prefixal) negation. It has considered the important status negation holds in the language, and indeed, as a basic characteristic of human behaviour and experience, and discussed the rationale for its choice as a case study. Finally, the corpus-based approach adopted by this research has been examined in respect of the characteristics
of such an approach, along with a summary of various attitudes taken by different researchers to corpus-based methodologies.

This thesis now re-examines and replicates the methodologies used in previous studies with a view to being able to provide more refined measures of both productivity, in general, and morphological negation, in particular, based on the use of authentic language data. The central argument of the thesis is that, while some earlier researchers utilised authentic language data, this research, by its use of the BNC, examines larger quantities of language data and does so in more detail. Although there has been previous research into both language productivity and prefixal negation, this investigation incorporates both these areas into one study, with the added advantage of a data-driven methodology. This, therefore, provides more contemporary and comprehensive coverage on the matter.

The thesis now continues with a review of the literature which relates to both productivity and morphological negation. Due to the chronological nature of the development of methodologies for the study of this topic, the literature, to some extent, is dealt with in a sequential manner.
Chapter 2

Literature Review

The introductory chapter examined the importance of language productivity; of negation, specifically morphological negation; and of the role of electronic corpora as a research tool. Negation, the use of affixation, and language productivity as a whole are well covered in the literature, and, of late, much of the work on productivity has involved the use of electronic corpora. Unlike dictionary-based accounts focussing on words which are potentially available for use, or models based on intuition, corpus approaches allow an insight into actual usage. Electronic corpus-based methodologies are still a relatively new phenomenon, but their influence is already significant in the field of lexicography and in a variety of areas of linguistic research.

There has been considerable research into the field of negation (Jespersen, 1917; Eifermann, 1961; Klima, 1964; Jackendoff, 1969: Smith, 1974; Tottie, 1980; Horn, 1989; Croft, 1991; Mazzon, 2004, etc.), though the majority of this work has investigated sentential negation (e.g. *She is not happy*), as opposed to morphological negation (e.g. *She is unhappy*). Similarly, much research has been carried out into the use of affixation (Aronoff, 1976; Bauer, 1992; Bauer, 1994; Bauer, 1996; Bolozky, 1999; Plag, 1999; etc.), particularly suffixation. However, little research has focussed on prefixation, and even less on negative prefixation. Individual aspects of the topic of this research have often been dealt with extensively, some even combining various aspects of this study. It is, however, highly doubtful as to whether any research to date has encompassed all the aspects to be covered by the current research, namely productivity; morphology; negation; prefixation; and utilising a
corpus-based approach, and it is this that defines the originality of this thesis. Due to the nature of the aspects covered, the current literature review discusses the various issues by means of both separate and combined examinations, dependent on the approaches adopted by the various researchers.

Thus, the review examines a variety of methods for determining productivity which have been employed over the last forty-five years. This incorporates the early studies, such as intuitive and dictionary-based methods, including the work of Zimmer (1964), Marchand (1969), Funk (1971) and Aronoff (1976). The review then continues by examining early corpus-based accounts, such as that undertaken by Tottie (1980), progressing to the classic corpus-based studies of Baayen, along with those carried out with his colleagues (Baayen and Lieber, 1991; Baayen, 1992; Baayen and Renouf, 1996). Although Kwon’s (1997) corpus-based study takes a historical perspective, this is also examined for its focus on the negative prefixes. Whilst occurring at various points in the chronology of methods, psycholinguistic approaches are also discussed, with particular reference to Baldi et al (1985), Frauenfelder and Schreuder (1992) and Schreuder and Baayen (1994). Finally, works are reviewed which utilise a combination of methodologies, for example, those by Plag (1999) and Bolozky (1999).

The current research provides a more sophisticated approach to measuring language productivity. Certainly, Baayen and his co-authors in the late 1980s and early 1990s took a massive and innovative step forward in terms of their corpus-based approach to this phenomenon, although their findings are not specifically reported in detail at this stage of this research. The reason for this is that the current study analyses their methods more closely in later chapters as they are being replicated. Initially, this is with a view to determining the productivity of the
negative prefixes of English, but more importantly, it also develops their methods further in order to produce a more fine-grained approach to measuring morphological productivity. Likewise, this study examines the variables which may also affect the productivity of a morphological process, such as semantic and historical factors, and the influence of word class.

2.1 Early studies in productivity - intuitive and dictionary-based approaches

With the advent of Chomsky’s generative grammar in the 1950s came a system which abstractly described what was and was not ‘allowed’ in a language. According to generative grammar, a language had a finite set of rules which, in turn, could generate an infinite number of potential constructions. Generative grammar became the zeitgeist, and inevitably attention began to turn to the output of those rules; productivity became the vogue of the 1960s and 1970s.

(T)he first who really tried to systematically use the concept of productivity as a criterion for the selection of those patterns that are relevant in Modern English word formation was Marchand (1951, 1955, 1960) [...]. It should be borne in mind, however, that Marchand developed his theory in a pregenerative framework [...] It was basically when generative grammars discovered word formation – the first full-scale treatment is Lees (1960) – that productivity really became a crucial problem.

(Kastovsky, 1986:587-588)

Prior to the 1980s, all approaches to productivity were, by necessity, based on intuition; perceived output; or on measures where the premise for evaluating productivity involved the number of inclusions of a particular process which were listed in a dictionary. Before the more widespread use of electronic databases, there was little alternative, though generally the pronouncements made were based on

Despite being written over forty years ago, Zimmer’s (1964) investigation into prefixal negation and language productivity still stands as a most prominent work in its field. In the introduction to his monograph, Zimmer sets out his aims as “an attempt to state the restrictions on the formation of adjectival antonym pairs by means of negative affixes” (1964:10), and Weinreich claims that Zimmer “succeed(s) in establishing some of the limits of generative grammar in derivational morphology” (Weinreich in the Foreword to Zimmer, 1964:4). Zimmer claimed that, regarding the description of a language, intuition remains an important tool, but concluded that experiments needed to be developed to test the theories relating to language productivity (93). Zimmer’s work, though exceptionally comprehensive in investigating productivity restrictions, pre-dates the extensive use of electronic corpora by some twenty years. Whilst stemming from the Chomskyan view of generative grammar, and despite Chomsky’s early views on the use of corpora (1962:159 in Aijmer and Altenberg (eds.), 1991:8), Zimmer acknowledged that linguists have accepted a far too flimsy explanation of productivity for far too long. Even at this stage, in 1964, the need for more empirical means of testing productivity was evident (1964:93).

Marchand proposed that “(w)ord-formation is that branch of the science of language which studies the patterns on which a language forms new lexical units” (1969:2). His aims were “not to give the fullest possible list of examples, but to offer a description of the trends of word-formation by a picture of the various formative types.” (1969:9). He did not seek to compile an exhaustive list of word formation types, since he acknowledged that this was an impossible task, stating that “new
words are coined every day” (1969:9). Nevertheless, in his study, Marchand (1969) investigated the phenomenon of productivity, albeit in a particularly broad sense, including compounding, suffixation and prefixation, in terms of word formation as a whole. His approach was both diachronic, charting the history of various elements of word formation, and categorical, in that it sought to classify word formation processes and individual examples into morphological, semantic and grammatical patterns. Marchand did not specifically limit the section of his study on prefixation to negative prefixes, neither did he engage overtly with productivity in the narrower sense, apart from providing brief reference to patterns or types which appear to have become more or less popular. In describing non-, for instance, he proposed that “The substantival type has become very productive recently” with examples such as “non-novel, non-resident, non-student.” (1969:180). Presumably, his findings are based on intuition to some degree, in addition to documenting negatively-prefixed words which he has encountered in speech or writing. Whilst extremely extensive in its scope, Marchand did not draw overall conclusions to his study, since its aim was to categorise “established patterns” of the language which were “representative of the structural system” (1969:9). Though pre-dating the use of corpora, Marchand’s research remains a significant work regarding the categorisation of the components of word formation and, thus, of productivity.

Two years later in 1971, Funk dealt specifically with negative affixes, though from a solely semantic perspective, by examining their synonymy. Funk’s aim was “to investigate the present-day results and consequences of the accumulation of negative affixes in the course of development of the English language.” (1971:364). He began with the premise that negative affixes might be considered “roughly synonymous morphemes”, but “that a language cannot for a long time afford such an
extravagance as [...] different formative elements denoting the same thing” and, as such, “we must be sceptical about the full synonymy of all negative prefixes in Modern English.” (1971:364). Funk’s study limited itself to the negation of adjectives, believing these to be “of primary importance within the field of affixal negation.” (1971:364). Funk did make judgements relating to the current productivity of the negating affixes, though these were founded on the number of inclusions in a variety of dictionaries and other non-finite, non-specific sources, from which he produced a semantic classification matrix:

<table>
<thead>
<tr>
<th>Table 2.1 Summary Matrix of Negative Prefixes after Funk (1971:374)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>semantic derivatives</strong></td>
</tr>
<tr>
<td>contrary meaning</td>
</tr>
<tr>
<td>‘is not self’</td>
</tr>
<tr>
<td>deverbial stems</td>
</tr>
<tr>
<td>‘has not been verb-ed’</td>
</tr>
<tr>
<td>deverbial stems</td>
</tr>
<tr>
<td>deverbial stems</td>
</tr>
<tr>
<td>deverbial stems</td>
</tr>
<tr>
<td>stylistic layer</td>
</tr>
<tr>
<td>literary</td>
</tr>
<tr>
<td>scientific</td>
</tr>
<tr>
<td>origin of stems</td>
</tr>
<tr>
<td>Romance</td>
</tr>
<tr>
<td>Neo-Latin</td>
</tr>
</tbody>
</table>

Funk determined that adjectives needing to convey strong evaluative meanings would tend to utilise un- in their negative form (unprofessional), whereas adjectives requiring purely descriptive characteristics would tend to use non- for their negative counterpart (non-professional). He concluded that the meaning of each
of the negative prefixes (except for non-\(^{-1}\)) varies, being determined by the semantic characteristics of the base words to which it is attached (1971:382). In other words, that it is semantic characteristics which determine productive output.

“Aronoff (1976), though still pre-dating the use of corpora, represents the first attempt in the literature on generative morphology to formalize the notion of productivity.” (Baayen and Lieber, 1991:802). Aronoff acknowledged the morphological restrictions on the base words to which affixes can be applied, but claimed that a relatively simple counting method could be employed to determine productivity, whereby the total number of ‘possible’ output words for a given process is set against the total number of ‘actual’ words (Aronoff, 1976:36). This ratio, he argued, could provide a simple measure of productivity for any word formation process.

There are obvious potential problems with Aronoff’s methodology, since quantifying both ‘possible’ and ‘actual’ words is to some extent subjective, and Aronoff himself recognised this. His method relied upon the notion that “every time we make up a new word, it is entered in a list. Unless all new words are listed, we have no effective procedure for computing the ratio of existing to possible words” (Aronoff, 1976:36). Although Aronoff’s method again preceded the use of electronic corpora, Baayen and Lieber (1991) subsequently examined and developed his proposals utilising a corpus to determine “a reasonable list of ‘actual’ words.” (803), as is discussed in Section 2.2 and also in more detail in Chapters 4 and 5.

Ultimately, Marchand and Funk took a semantic stance, rather than one which is specifically productivity-focussed. They considered and categorised word formation elements, based on intuitive comment and historical data from

---

\(^{-1}\) “It \([non-]\) is open to free use for strictly contradictory negation (which simply means that the adjective does not apply).” (Funk, 1971:282).
dictionaries. Both Zimmer and Aronoff however, specifically considered productivity, albeit from the perspective of base-word restrictions, and recognised the need for more empirical quantitative measures. Thus, research into productivity began in earnest as a natural consequence of the generative approach to rule-based word formation. The methodologies for measuring productivity employed in the 1960s and 1970s were inaugural exploratory techniques, which examined the phenomenon from a different perspective than had previously been tried, and these methods were facilitated by the emergence of the generative era. The means of determining productivity at this time were putative and based on speculative theories, and, unless dictionary listings were taken into consideration, no empirical quantitative measures were used.

Thus, the intuitive and dictionary-based approaches lay the foundations for the investigation of subsequent hypotheses relating to productivity. Corpus-based measures would soon allow intuitive and dictionary-based theories to be tested empirically for the first time.

2.2 Corpus-based approaches

This section deals with the development of the corpus-based methodologies which began to be applied in the early 1980s. As a result of these developments, earlier studies involving intuitive and dictionary-based theories began to be revisited and replicated with the use of more empirical measures. These methods were being developed in response to the need to quantify productive output, which was facilitated by advancing computer technology. From the mid-1980s, the more widespread use of electronic corpora began to filter through, although it is from the early 1990s that there was a significant rise in their availability and subsequent use.
As a result, corpus-based methodologies were developed, and large scale studies involving the ‘actual’ language produced by speakers and writers became possible.

Tottie’s 1980 study of negation was among the first to utilise corpus-data. She acknowledged that most previous research into affixal negation had been concerned with the productivity of the various negative affixes, or alternatively, had examined the synonymy of affixal negation as compared with sentential negation (101). Tottie’s aim was to use a corpus to investigate the distribution of affixal and non-affixal negation, along with an examination of the constraints on this distribution (1980:102). Though small by today’s standards, the corpus, from the Survey of English Usage at University College London, consisted of 100,000 words, which was approximately equally divided between written and spoken texts (Tottie, 1980:104).

Tottie examined the occurrence of all the negation in the corpus, which consisted of 181 examples of affixal negation and 180 examples of non-affixal negation (1980:104). She then went on to manually investigate the individual occurrences in terms of the various constraints which make affixal and non-affixal negation obligatory (1980:105). Her results demonstrated that there are a larger number of factors which dictate the use of affixal negation than non-affixal negation, and that typically there are more constraints on written rather than spoken language (1980:114). It was found that the most important factor preventing the use of affixal negation, and thus dictating the use of non-affixal negation, was “the lack of AF [affixal negation] counterparts to many adjectival” (1980:109). Despite the relatively small corpus, Tottie was also able to demonstrate that, in 10% of cases, affixal and non-affixal negation were interchangeable, but that in the majority of cases, “(t)he existence of both affixal and non-affixal negation in English is thus not a freakish
redundancy of the language but the variants occur in almost complementary distribution, to meet different needs” (1980:122).

It was the 1990s which saw a huge rise in the use of electronic corpora to investigate language productivity, with studies such as Baayen and Lieber, 1991; Baayen, 1992; Frauenfelder and Schreuder, 1992; Schreuder and Baayen, 1994; Baayen and Renouf, 1996; Kwon, 1997; Plag, 1999; and Plag et al, 1999. All the above research examined productivity, although none, except Kwon, specifically in terms of negative prefixation. As can be seen from this series of corpus researchers, Harald Baayen figures significantly in studies relating to productivity at this time.

The broad issue of derivational affixes was investigated by Baayen and Lieber in 1991 in order to compare statistical measures against intuitive notions of productivity. Baayen and Lieber adopt a definition of productivity proposed by Schultink in 1961 that does not involve intentional coining, such as clippings, blends, etc., and which also, therefore, distinguishes productivity from creativity. “By productivity as a morphological phenomenon we understand the possibility for language users to coin, unintentionally, a number of formations which are in principle uncountable….” (Schultink, 1961, translated by Van Marle, 1985:45). However, Baayen and Lieber claim that if an affix is genuinely productive, any subsequent complex word formed using that affix “will go unnoticed” by the hearer/reader, and that likewise, when an occasional new word is coined using an unproductive affix, “such coinages will always draw attention to themselves.” (1991:808).

The aim of Baayen and Lieber’s study is to “review some of the definitions and measures of productivity discussed in the generative and pregenerative literature”, and to propose a new measure “based on the token frequencies of types”
A ‘type’ is a specific word form which occurs in a corpus, and a ‘token’ is the number of times that specific type occurs in the corpus as a whole. For example, the adjective *unlawful* is an example of a ‘type’ in the BNC, which occurs 902 times in the corpus as a whole; thus demonstrating a ‘token’ frequency of 902.

Baayen and Lieber also introduce the notion of hapax legomena, which are words that occur only once in a given corpus, and specifically they claim that a productivity measure which is based “on the number of hapax legomena for a given affix in a corpus” produces results which concur “with our intuitions about productivity.”

Based on their adoption of a definition of productivity proposed by Schultink (1961), they propose that a higher mean token frequency of an affix correlates with a lower degree of productivity and vice versa, a lower mean token frequency indicates a higher level of productivity (1991:801 and 809). They therefore propose that the number of types for a particular affix with a frequency of one, divided by the total number of tokens of words with that particular affix indicates the level of productivity for that affix. The introduction of hapax legomena as a relevant component in the measuring of productivity was seen as a major breakthrough in the early 1990s studies and forms a key part of the research presented in later chapters of this thesis.

Baayen and Lieber (1991) concur with Schultink (1961) in stating that productive affixes are, in theory, capable of producing an unlimited number of new words (neologisms), whereas unproductive affixes will coin a limited, finite number of new words. As with Schultink, Baayen and Lieber’s methods take into account the potential restrictions (phonological, syntactic, semantic) which might prevent certain affixes attaching to certain base words. As a result, they accept the notion of differing degrees of productivity (Baayen and Lieber, 1991:801). Their research
demonstrates that the base word that an affix may be attached to has “an unexpectedly strong effect on the degree of productivity of derived words”, and that the development of their methodology allows for a distinction to be made between truly productive affixes and those which are “sporadically found in neologisms” (839). Baayen and Lieber’s (1991) findings suggest that, having applied a “mathematically motivated measure of morphological productivity”, they are able to demonstrate “a satisfying ranking of affixal classes according to their degree of productivity” (838-839). Moreover, it is their belief that:

....the gap between theoretical linguistics on the one hand and quantitative linguistics on the other is not unbridgeable, and that the interaction of the two approaches aids the goal of understanding the phenomenon of human language.

(1991:839)

The degree to which these aims were met are discussed in further detail in the following chapter, where Baayen and Lieber’s (1991) measures are replicated using a larger corpus, and their findings are evaluated more closely.

Following on from Baayen and Lieber’s 1991 study, Baayen (1992) sought to examine a number of productivity measures which were based on quantitative approaches to “show that productivity and frequency are indeed closely correlated.” (111). Here, Baayen was again advocating the usefulness of hapax legomena in determining the productivity of an affix. Baayen’s overall aim was “to clarify some of the issues involved in the quantification of morphological productivity.” (1992:110).

Baayen’s view of the usefulness of hapaxes and the relationship between productivity and frequency is, however, questioned by Van Marle, also in 1992,
when he argues that once a word is coined, its subsequent frequency of use is of little relevance to productivity (156). Van Marle’s paper is a direct comment on Baayen’s (1992) study, and according to Van Marle, the accepted interpretation of morphological productivity is the ability of processes which enable them to produce neologisms. Whilst disagreeing with Baayen on a number of issues, Van Marle, nevertheless, states that:

(i) in my opinion there can be no doubt that Baayen’s work on morphological productivity [...] is both highly provocative and important. No one seriously involved in the study of morphological productivity can afford to leave Baayen’s work unread (although it does not make easy reading due to the sophisticated statistical and mathematical procedures employed). [...] The overall importance of Baayen’s work is undisputable and it seems unquestionable that his work will greatly deepen our understanding of some of the basic properties of the phenomenon of morphological productivity. (1992:152)

Van Marle believed that Baayen’s approach to productivity required more precision for a number of reasons. (1992:151). Firstly, simply defining the concept of a new word is problematic. For example, at what point is a new word said to exist? Is this at the moment the word is first coined, and how would this be recorded and verified? Is it as a word begins to become established, or is it the point at which it is recognised by lexicographers for inclusion in a dictionary? Van Marle also asks the question “how many ‘new’ words should a morphological process be able to produce in order to be called productive?” (1992:151). Van Marle further questions why not all affixes can be considered productive, and of those which do possess this quality, “why not [all] to the same degree?” (1992:151). He noted that most morphological research had, up to Baayen’s involvement, focussed on the language system, “whereas the performance dimension of morphological productivity had to a large extent remained un(der)-explored.” (1992:151). He went on to acknowledge
Baayen’s contribution in terms of the correlation between frequency and productivity, and Baayen’s view that frequency has an extremely significant part to play in any measurement of language production, proposing that in Baayen’s research “performance factors play a much more prominent role” than in earlier studies (1992:152). However, Van Marle disagrees with Baayen on a number of issues. For example, he claims that Baayen is more concerned with the degree to which a process is likely to coin new words, as opposed to its ‘ability’ to create new coinages (1992:152). Similarly, Van Marle claims that some of Baayen’s findings appear to be counter-intuitive. For example, Van Marle points out that typically “it is the unmarked categories which are considered the most productive” (1992:154). In Baayen’s 1989 study, however, the findings suggest that of the Dutch ‘personal names’ suffixes -er (neutral and unmarked version) and -ster (female and marked version), “the marked category in -ster ranks 3½ times higher than that of its unmarked counterpart in -er!” (Van Marle, 1992:154). More significantly for this study, however, Van Marle argues that the notion of hapaxes is “a too simplistic conception of the relationship between productivity and frequency” (1992:161).

Much of Baayen’s methodology relied heavily on the value of hapaxes in determining the presence of neologisms and, in turn, in measuring productivity. However, Van Marle (1992) “calls into question Baayen’s general starting-point according to which hapaxes are always considered indicators of productivity” (157). Van Marle argued that “many hapaxes do not have anything to do with morphological productivity […]”, noting that “many complex words that are attested only once […] are perfectly common, actual words” (Van Marle, 1992:157). Van Marle did not necessarily reject the role of hapaxes entirely, but doubted Baayen’s tendency “to equate hapaxes with newly coined formations (and,
consequently, to consider them indicators of productivity)” (1992:157). Instead, Van Marle (1992) submitted that the notion of hapaxes required further, more sophisticated investigation, along with the use of larger databases than Baayen worked with at this time. As with other studies, Baayen (1992) and Van Marle (1992) did not concentrate either on prefixation or negation. Nevertheless, Baayen in particular, was perhaps the most instrumental and influential in advancing the role of corpus-based methodologies at this time, which in turn allowed others to apply his formulae and develop his methods to suit their specific field of interest.

Derivational affixes were again the focus for Baayen and Renouf’s 1996 study in which several measures of productivity were compared, based on the occurrence of neologisms in newspaper data. Though they focussed primarily on derivational affixes, Baayen and Renouf’s study did include two negative prefixes, namely “the rival prefixes un- and in-” (1996:70). As with Baayen and Lieber’s (1991) study, Baayen and Renouf (1996) emphasized the importance of hapax legomena in any measure of affixal productivity. They proposed that unproductive affixes are characterised by a large number of types with high token frequencies. Conversely, Baayen and Renouf propose that unproductive affixes are typically distinguished by a particularly low proportion of types with low token frequencies, and hapax legomena are especially rare in the case of non-productive affixes. Indeed, Baayen and Renouf went so far as to state that types with “large numbers of hapax legomena are a sure sign that an affix is productive.” (1996:74). They did go on to acknowledge that “(t)he hapax legomena in a sample need not be neologisms” (1996:74), but still concluded that “(n)eologisms are significantly overrepresented among the hapax legomena.” (1996:76).
Baayen and Renouf’s overall findings was that the prefix in- is barely productive (1996:86), whereas un- was considered to be highly productive (93). The findings of this study suggested, just as Funk proposed in 1971, that it is the morphological structure of the base word, rather than the affix itself, which is of significance and which determines productivity. For example, -ly freely attaches to adjectives to create the corresponding adverb (typical – adjective >> typically – adverb; quick – adjective >> quickly – adverb), but is restricted from attaching to adjectives which already end in -ly (elderly >> *elderlily; worldly >> *worldlily). Baayen and Renouf referred to these as “ephemeral formations”, proposing that examples such as sillily and friendlily “are exceptions to a rule rather than the result of a rule with a very low degree of productivity” (1996:83), though the OED does list sillily and friendlily (as cited in Bauer, 1983:89). Likewise, Baayen and Renouf claimed that the highest level of productivity for adjectival un- formations were those with base words in -ed, for example unnannied, and that verbal formations in un- were rare and that, therefore, un- was unproductive in terms of its use in verbs (1996:85). Thus, their study also proposed that there are structural constraints on the affix as to the base word to which the affix may attach.

Both the Baayen and Lieber 1991 study and the Baayen and Renouf 1996 study are ground-breaking in their methodology and application; using much larger-scale corpora than had previously been used; and involving the occurrence of hapax legomena as an important factor in determining productivity. Neither of these studies deals conspicuously with negative prefixes, though both studies incorporate the prefixes in- and un-, along with a variety of other derivational affixes, both prefixes and suffixes.
Kwon’s doctoral thesis (1997) specifically examines negative prefixation. The approach, though corpus-based, adopts a diachronic perspective, charting the emergence and use of the various negating prefixes of English from 1500 to the present day. Kwon’s thesis investigates the co-existence of five negative prefixes: *a-*, *dis-*, *in-*, *non-* and *un-*, though the rationale for Kwon’s choice of prefix is not made explicit. In addition to utilising corpus-based data, the study also makes use of dictionary-based material, proposing that “the information contained in dictionaries provides an account of perceived correct usage, while computer-based exploration of corpus texts gives access to actual usage” (Kwon, 1997: Synopsis). The study does not aim to engage specifically with the notion of productivity, its scope being deliberately restricted to an account of the diachronic use of the various negating prefixes, which may, nevertheless, prove invaluable in the development of synchronic semantic and grammatical profiles for each of the prefixes. Kwon proposes that it is important to examine past and present usage in order to be able to predict the future (1997:229). He proposes that “(t)he frequency and collocation information extracted by this analysis, together with a comparison with historical data, provide a useful basis for assessing the direction of the future development in negative prefixation.” (1997:230).

Similarly, the frequency with which forms of existing morphological patterns occur may contribute significantly to establishing the productive status of the processes. Familiarity with the typical morphological patterns of negative prefixation is also important for producing less obvious words, or new words. (Kwon, 1997:231)

Unlike many other studies (Marchand, 1969; Algeo, 1971; Funk, 1971; Bauer, 1983; Hulse, 2002; Mazzon, 2004; Hamawand, 2009), however, Kwon’s findings suggest that “there has been no increase in the proportion of *non-* words
over the past three decades or so [....] although the prefix *non-* exhibits a considerable degree of potential for combining with a wide range of bases” (1997:231). Kwon also submits that the same could be said of the prefix *un-* although “some established patterns such as ‘un-X-able’ and ‘un-X-ed’ would still be used as the basis for creating new words.” (1997:231). In conclusion, Kwon proposes the need for a monitor corpus, which continues to develop “with ever-moving stores of text” and this, he suggests, will provide a corpus with “a historical dimension with detailed evidence of language evolution.” (1997:232).

Corpora, though still considered a relatively new phenomenon, to some extent, have now been around for over thirty years; constantly developing and becoming more sophisticated tools with which to analyse language. While the 1980s had seen the start of the early use of corpus-based material, at this stage, the scale of such corpora was negligible in comparison with those available today. Likewise, related software and the tools integrated into corpora, which were used to extract and exploit the appropriate data, were limited, somewhat primitive and unreliable by today’s standards.

Until the mid-1980s corpus linguistics typically involved mainframe computing [....], it could take an hour or more to make a concordance [....] In the early 1980s a captive computer scientist or friendly computer programmer was almost indispensable to assist many aspiring corpus linguists to cope with inevitable technical problems associated with data management and the programming skills necessary for corpus analysis.

(Kennedy, 1998:7)

By the 1990s, however, huge projects were being undertaken to compile new corpora, such as the British National Corpus, which were more reliable and vastly superior in terms of how the data could be manipulated and utilised. Thus, as empirical quantitative measures have developed and electronic corpora have become
larger and more sophisticated, so the methodologies relating to productivity have also become more refined and more reliable. The studies by Baayen and Lieber (1991), Baayen (1992), Baayen and Renouf (1996), and Kwon (1997) have all benefited from the use of corpora and have been able to demonstrate the usefulness of quantifying output in terms of productive usage by means of capturing attested language.

The current study further examines Baayen’s theory that *token* frequency is an important factor in a measure of productivity, and also investigates the usefulness of hapaxes to determine the extent to which they can be considered neologisms. In addition, this research also examines Van Marle’s (1992) view that there is a correlation between hapaxes and *types*, and that type frequency may indeed be a useful variable in measuring productivity (158). In utilising the BNC, this research makes use of a much larger corpus than was available to Baayen and his co-authors in the 1990s.

### 2.3 Psycholinguistic approaches

Psycholinguistic approaches to productivity are taken by Baldi et al in 1985, Frauenfelder and Schreuder in 1992, and by Schreuder and Baayen in 1994. Baldi et al’s (1985) approach is purely psycholinguistic, with no recourse to corpora. Frauenfelder and Schreuder’s (1992) research, though psycholinguistic in nature, examines Baayen’s approach, believing it has much to offer, but that it requires further development. Schreuder and Baayen’s (1994) study was one of the earliest psycholinguistic enquiries to utilise corpus data.

Negative adjectives were the focus of Baldi et al’s (1985) research; the question asked being: “Is it possible to demonstrate experimentally which negative
prefix or prefixes are the most productive?” (34). Their methodology followed a psycholinguistic approach “designed to test native speakers’ intuitions about negative prefixes” (1985:36). In the first experiment (a production task), the participants were asked to lexically negate newly-introduced words for which there were no pre-existing negative forms. In their second experiment (a preference task), participants were given a series of both genuinely-negated and pseudo-negated adjectives and, in a multiple choice task, they were asked to select their preferred negated form. It should be noted, however, that all the negated forms chosen for the second experiment were prefixed with in- (and its variants), non- and un-. These prefixes were selected for the second experiment as they had been the most common responses in the earlier production task. Baldi et al’s experimentation was stringently carried out, with a great deal of foresight as to the nature of the input. For instance, both Latinate and native pseudo-words were included, and the number of syllables and the type of suffixes the words contained were taken into account, based on the Latinate or native nature of the pseudo-word. Latinate and native spelling and consonant peculiarities were also incorporated into the pseudo-words, and Latin phonological constraints were also taken into consideration.

Overall, it is unclear whether Baldi et al answered their research question in terms of the productivity of the negative prefixes. Certainly, they demonstrated that the prefix un- is chosen more often in their experiments than any other negative prefix, especially with native-type pseudo-words, though their claim that their findings “show quite clearly that un- is indeed the most productive negativizing prefix in English today” is open to doubt (1985:54). Their findings also suggested that “(n)on- is the strongest second choice”, though this may be due to “the fact that it never assimilates phonologically to the following segment” and that it “readily
attaches to all roots, regardless of etymology.” (1985:54). The overall question remains, however, as to whether ‘greater preference’ and ‘chosen more often’ equate to productivity, since the input incorporated pseudo-words and, furthermore, was limited to pseudo-adjectives. Likewise, the participants of the research were linguistics/psychology university students, who, it could be argued, might have been expected to have a greater-than-average knowledge of the language and its etymology.

Baldi et al’s (1985) study focused on the transparency of the components in a morphologically-complex word, as well as the frequency of the base word and its affixes, as significant factors in determining productivity. Frauenfelder and Schreuder (1992), on the other hand, did not specifically consider prefixes or suffixes, but more broadly examined the implications of productivity for lexical processing and the storage of complex words, which they felt had been much ignored by psycholinguists. Frauenfelder and Schreuder acknowledged Baayen’s contribution to “reinstate the issue of morphological productivity on the psycholinguistic agenda” (1992:165), proposing he achieved this by developing a quantitative measure of productivity based on word token frequency (181). Baayen’s model assumed that morphological productivity plays an important role in determining how a complex word is processed. Frauenfelder and Schreuder referred to his model as a ‘race’ model, with two processing routes operating in parallel (1992:172). One processing route, storing morphologically complex words as single units (the direct route) is motivated by the principle of the ‘economy of processing’. The second processing route, decomposing morphologically-complex words into bases and affixes (the parsing route) is motivated by the principle of the ‘economy of
storage’ (1992:181). For Frauenfelder and Schreuder, whichever processing route achieves word recognition first, is the winner of the ‘race’.

Frauenfelder and Schreuder (1992) found that “the direct access route presumably wins the race for most English words.” (180), but they also examined other languages with highly productive morphologies, such as Turkish, which is also phonologically transparent. Their findings suggested that a complex word may in fact be recognised by a combined effort between the direct and the parsing route. As a result, Frauenfelder and Schreuder extended Baayen’s model, proposing a link between productivity and the phonological and semantic transparency of the word forms (1992:182). They argued that the parsing route is most likely to win for transparent, low frequency words, typically containing highly productive affixes. Alternatively, the direct route would win for high frequency opaque words (typically those which are difficult for parsing) containing unproductive affixes (1992:182).

Schreuder and Baayen (1994) examined prefixes (though not exclusively the negating variety) in both English and Dutch, again investigating the psycholinguistic processing of morphologically-complex words. Once again, this study did not overtly investigate productivity; instead it used corpus data in relation to token frequency, to ascertain whether the components of morphologically-complex words are involved in lexical access. In the 1970s and 1980s, Taft et al had proposed and developed a model of ‘prefix stripping’:

According to this model, a prefixed word is stored in the lexicon as a representation of its stem, with information stored within this lexicon entry about what can combine with the stem to form a word. Thus when a prefixed word is to be recognized, it is firstly decomposed into its prefix and stem, and lexical access then proceeds on the basis of the stem only. Once the appropriate lexical entry is located on this basis, the prefix information stored within this lexical entry can be consulted and the word can thus be recognized.

(Taft, 1981:289)
Taft and Forster (1975) proposed that prefixed words are analysed into their component parts in order for lexical access to occur. Their experiment involved decision tasks, whereby subjects were presented with a list of words and had to decide whether they were genuinely-prefixed words or not (639). Their findings suggested that complex non-words which incorporated a genuine base word and a genuine prefix took longer to process than non-words which did not include a genuine base and prefix (640). Schreuder and Baayen (1994) sought to evaluate what had become, in the 1970s and 1980s, the influential ‘prefix stripping’ model, though their study utilised corpora in order to demonstrate that “prefix stripping does not enhance processing efficiency […] as originally claimed by Taft and Forster [1975]” (1994:357). Schreuder and Baayen, however, concluded that it is “highly improbable that prefix-stripping […] is involved in lexical processing” (1994:373).

Though Baldi et al (1985) did not utilise a corpus approach, the psycholinguistic nature of their study allowed an insight into the considerations that a speaker may take into account in the production of a neologism. Their study, though pre-dating the majority of corpus-based research, therefore stands as a useful precursor to corpus-based methodologies. Both the Frauenfelder and Schreuder (1992) and the Schreuder and Baayen (1994) studies utilise Baayen’s quantitative definition of productivity, based on word token frequency. This methodology is discussed in much greater detail in the next chapter, and is replicated using a larger corpus than was available to earlier studies, to provide “an objective, statistical measure of productivity” for each of the five negative prefixes (Frauenfelder and Schreuder, 1992:171).
2.4 Combination approaches

As corpus-based methodologies began to become more widespread, some researchers, whilst acknowledging the usefulness of quantitative measures, also began to employ a variety of different measures to test productivity. These can be referred to as ‘combination approaches’. Notable advocates of these types of applications since the 1990s are Ingo Plag and Shmuel Bolozy.

Plag (1999) sought to bridge the gap between theoretical and empirical approaches to productivity, focussing on derivational suffixes, especially verb-forming affixes such as -ize, -ify, and -ate. He argues that the rules governing word formation should be understood in terms of being base-driven, rather than affix-driven and takes the position that productivity is a derived property of language, which is the result of other mechanisms. In other words, Plag argues that productivity is an epiphenomenon, or a secondary feature of language, as opposed to being an inherent property of morphological rules (1999:12). He examines whether productivity should be viewed as a quantitative or qualitative notion, but argues that the two approaches are not mutually exclusive (226). Plag’s main focus is on a qualitative approach, and he proposes that the productivity of an affix can be typically “predicted on the basis of the process’s peculiar structural properties and restrictions.” (1999:244). However, he also conducts a quantitative investigation of the productivity of verb-forming affixes, using dictionary data (chiefly from the Oxford English Dictionary), in addition to corpus data, believing that a combination of the two sources can help eliminate the bias inherent in either source. Plag argues for a model of derivational morphology which involves the relationship between form and meaning in complex words, in other words, a model based on the output of morphological processes (1999:234), which predicts the approximate synonymy of
affixes (240). Plag identifies the structural constraints on the affixes involved as the most significant factors in governing productivity, and further determines that the discourse type has a significant influence on the productivity of affixes. His output model of derivational morphology argues for the notion that “(t)oken-blocking (but not type-blocking) is certainly a relevant factor restricting productivity” (1999:243), and goes on to propose that:

....the OED can be used even for measuring the productivity of a given process, provided that one is aware of its shortcomings. The application of Baayen’s text-corpus-based measures to derived verbs as manifested in the Cobuild corpus indicated that these statistical measures are useful for the determination of the productivity of overt affixation, if employed carefully, and if complemented by a qualitative investigation. [...] The foregoing study can therefore be read as a case study laying out a possible research agenda for future in-depth investigations of other derivational processes in English and other languages.

(Plag, 1999:244)

Bolozky’s 1999 research presents three separate means of evaluating productivity, proposing that “no single method constitutes a sure criterion of productivity on its own” and that only “when findings based on different methodologies all converge and point to it as such” should a process be considered as truly productive (1999:3). His three methods are productivity tests involving open-ended and judgement tasks; dictionary-based comparisons; and corpus-based methods involving the ratio of hapax legomena to token frequency. Like other researchers, Bolozky proposes that in terms of token frequency “for a variety of reasons, there exists some correlation between low frequency and high productivity.” (1999:6). Again, the study focuses primarily on derivational suffixes, though there is cursory mention of prefixation. Bolozky acknowledges that all his methods, apart
from the dictionary comparison, require improvement; the corpus method, for example, needing “larger and more varied corpora than are currently available” (191).

As with Plag (1999), Bolozky (1999) argues that productivity is base-word-dependent, as opposed to being affix-driven, and he maintains that, if conducted carefully and adequately, dictionary comparisons “can be quite reliable, and eventually very useful.” (194). Bolozky concludes that the relative productivity of a derivational affix may be calculated by the application of two different ratios. Firstly, he proposes that in a lexicon, dictionary or corpus, the ratio of the number of neologisms compared to the total number of tokens within that particular category is a useful measure of productivity. For example, the number of neologisms created using the negative prefix *in*- could be calculated and this would then be compared to the total number of *in*- tokens in the whole of the particular corpus. The higher the ratio, the more productive the prefix would be considered to be. The second ratio Bolozky proposes is comparing the number of neologisms to the total number of neologisms in the same semantic category. So, for instance, by this measure, he is proposing that the total number of *in*- prefixed neologisms should be calculated and then compared to the total number of negatively-prefixed neologisms in the whole corpus (i.e. all neologisms created using any of the negative prefixes). Bolozky proposes that it is the second ratio which ultimately determines productivity, but that when the two ratios concur, this provides a more reliable measure (194). However, as referred to earlier, determining what constitutes a neologism is not always straightforward, and quantifying them presents unique problems; these issues are taken up in subsequent chapters.

---

2 Bolozky’s research focuses on Modern Hebrew, specifically Israeli Hebrew, and it is to this that he refers when he proposes that corpora need to be larger and more varied.
In 2001, Bauer undertook the task of presenting a coherent view of productivity from an eclectic standpoint, gathering from a vast array of sources and opinions. At the start of his research, he quotes:

‘Produktivität’ zählt zu den unklarsten Begriffen der Linguistik.
‘Productivity is among the least clear concepts in linguistics.’


Bauer goes on to suggest that productivity “may mean different things to different people” and that even attempting to provide a definition of productivity is “a matter of some dispute.” (2001:1). Throughout his book, he attempts to develop a definition of productivity, despite the differing views on the topic. He identifies that for some productivity is a “yes/no matter”, whereas for others it is “a matter of gradience” (2001:10). Ultimately, Bauer divides productivity into two discrete ideas; those of ‘availability’ and ‘profitability’, whereby ‘availability’ refers to whether “a morphological process is available or it is not.” (2001:205). On the other hand, ‘profitability’ refers to the extent to which an available process is actually used (2001:207). Bauer’s final definition of productivity is as follows:

‘Productivity’ deals with the number of new words that can be coined using a particular morphological process, and is ambiguous between the sense ‘availability’ and the sense ‘profitability’. The availability of a morphological process is its potential for repetitive rule-governed morphological coinig [....] determined by the language system [...]. The profitability of a morphological process reflects the extent to which its availability is exploited in language use, and may be subject unpredictably to extra-systemic factors.

(2001:211)

---

1 Bauer’s translation.
Bauer’s overall aim is “to reconcile the conflicting notions” relating to productivity, since, up to the time of his writing, there had “not been a coherent picture presented of what it is and what its implications are.” (2001:222-223).

2.5 Literature summary

The literature discussed here has covered an array of distinct, though often related, issues. Some studies have combined different approaches, but none to date have tackled synchronic productivity in terms of negative prefixation, from a strictly corpus-based perspective.

In the early days of generative grammar, the productivity of morphological rules began to come to the fore. Initially, and by necessity, the methods used focussed on intuition and dictionary-based approaches, though researchers began to acknowledge the need for more empirical methods, and as electronic corpora developed, so these were utilised to test existing theories of productivity. Generally, the new methodology complemented and supported previously held beliefs concerning productivity, but again, the need to develop the corpus method further into a more refined tool was recognised. Gradually, this allowed access to larger corpora and intuitive ideas and dictionary-based methods were able to be tested, along with psycholinguistic and combination approaches.

Broadly speaking, this study’s starting point is similar to Baayen’s - that statistical, quantitative and empirical methods provide the best means of assessing morphological productivity. It, therefore, applies a hapax-based approach to negative prefixes. However, the current study builds heavily on Van Marle’s criticism of Baayen’s approach, in attempting to fine-tune the methodology, by closely assessing Baayen’s model for reliability and suitability, partly through the introduction of a
qualitative dimension to the research and partly through a comparison of a variety of measures.

The next chapter firstly discusses the rationale for both the corpus approach and the choice of corpus. The chapter also determines and justifies which negative prefixes are to be assessed in this study and, once this has been established, the chapter goes on to examine the creation of the database of negatively-prefixed words from the BNC. Finally, the historical and semantic factors which may affect the negative prefixes are also investigated.
Chapter 3

Data Collection: developing the database and examining the prefixes

As discussed in Chapter 1, this research investigates the phenomenon of morphological productivity, and particularly engages with the controversial issue of how language productivity is measured. Thus, the study re-examines and re-creates earlier methods of measuring productivity proposed by researchers over the past forty years, with the intention of making use of previously-proposed formulae to create a more fine-tuned measurement system. By using an electronic corpus as the basis for its data gathering, the current research takes the negative prefixes of English as its case study, in order to optimize previous methodologies.

It has been established that productivity is a vital feature of language, yet the notion of productivity remains contentious. This research intends to advance the concept in order to help clarify the theories which relate to productivity, by gathering and investigating large quantities of relevant examples to initially provide a quantitative analysis, with subsequent qualitative evaluation based on this data.

The current chapter reports on a variety of searches conducted on an electronic corpus in order to create a database which could then be used as the basis for the analysis. The analysis of this data, to a large extent, replicates the methodologies proposed by others, such as Plag (1999), Bolozky (1999), Baayen and Renouf (1996) and particularly Baayen and Lieber (1991), though it also scrutinises and critiques these methods and to develop them further. The methods employed are
typically of the automated type but, as the research also examines the current limitations of such methodologies, some aspects also adopt certain manual examinations, editing techniques and analyses. The procedures to meet the aims of this study involve the selection of a relevant corpus; finding and extracting all of the negatively-prefixed words in that corpus; and the creation of a searchable database based of these negatively-prefixed words.

Once the negating prefixes to be used in this study have been selected, and the database created, an outline of the historical and semantic factors which affect those prefixes will be presented. Although this thesis is primarily about methods of measuring morphological productivity, there are of course many, often complex, and inter-related reasons why one prefix may have come to be more (or less) productive than another. The final sections of this chapter, therefore, explain some of these reasons, and focus particularly on the etymology and semantic differences of the prefixes under scrutiny.

3.1 Why a corpus-based approach?

The historical approach to productivity was, by necessity, based on intuition. Invariably, however, productivity was treated in slightly different ways by different researchers, and there was, and is, no definitive description of what constitutes productivity. As cited at the outset of this thesis, Crystal’s *Dictionary of Linguistics and Phonetics* defines productivity in the more restricted sense as “the use made by a language of a specific feature or pattern. A pattern is productive if it is repeatedly used […] to produce further instances of the same type” (1997a:310). Similarly, Matthews’ *Concise Oxford Dictionary of Linguistics* suggests that, in broad terms, productivity is “The property of permitting novel combinations of elements”, but
more precisely, refers to the specific rules which relate to constructions and formations, for example, “the formation of adjectives in –able is productive in that speakers can readily create new ones” (1997:297). Without a clear definition of productivity, it would seem problematic in the extreme, therefore, to utilise an approach based purely on intuitive notions. As Aronoff argues “The term productivity is widely used in studies of derivational morphology, and there is obviously some intuition behind the usage, but most of the discussion of it is rather vague.” (1976:35). Clearly, intuitive perspectives do not lend themselves to quantitative approaches.

As shown in the previous chapter, intuitive notions of productivity have also been used in conjunction with dictionary-based accounts, though many researchers, such as Aronoff (1976), Lieber (1992), Clark (1993), and Baayen and Renouf (1996), have discounted the use of data from dictionaries, as they believe they are not a dependable measure of productivity. They argue that dictionaries fail to include the most likely and regular complex forms (not every word prefixed with un- or suffixed with -able is listed, for instance), and that dictionaries lack in their ability to distinguish between currently and formerly productive patterns. Few, apart from Bolozky (1999), still advocate the use of dictionaries as an aid to determining productivity. With regard to the material included in a dictionary, by necessity this reflects the personal preferences of the lexicographers concerned, and serves to represent ‘actual’ significantly-attested words. Neologisms, rare words and full listings for all complex words (affixed and compounded) may be omitted.

A corpus approach, on the other hand, is based on more than intuition; on more than the views of the lexicographer. It allows for ‘genuine’ examples, including neologisms, relatively infrequent words, and simplex and complex words to be taken
into consideration. In recent years, electronic corpora have been progressively employed to equip a variety of language models with empirical evidence of authentic language use.

Linguists have always needed sources of evidence for theories about the nature, elements, structure and functions of language, and as a basis for stating what is possible in a language. At various times, such evidence has come from intuition or introspection, from experimentation or elicitation, and from descriptions based on observations of occurrence in spoken or written texts. In the case of corpus-based research, the evidence is derived directly from texts.

(Kennedy, 1998:7)

Thus, corpora can be used as evidence to test theories and to evaluate generalisations relating to language. They serve to substantiate facts about language and can contribute underpinning support for (or alternatively controvert) linguistic theories. In other words, corpora can lend support to the lexicographer’s judgement and corroborate intuition, and indeed, almost all dictionaries of note are now based on the use of corpora. As Jones suggests, when he describes intuitive approaches, “….these intuitions are often sophisticated and usually reflect a wider consensus. However, corpus data helps to eliminate the elements of chance by tapping into not one mental lexicon, but thousands.” (2002:21). There is no completely theory-driven linguistics and no completely data-driven linguistics (Leech, 2004); the use of electronic corpora can, therefore, provide useful data with which to test the theories. By the use of corpora, generalisations can be made about language, and suggested explanations provided.

Naturally, corpus-based studies demonstrate the ‘observed’ construction, rather than the ‘permitted’ construction (Bauer’s ‘profitability’ and ‘availability’ respectively, 2001) and are, therefore, used to evaluate the way in which language is
used in authentic contexts. However, corpora are now also used to help provide
descriptions of the grammar of a language. This does not, however, imply that
theoretical issues are not also the concern of corpus linguists. Indeed, “Corpus
linguists assume that relevant theories or hypotheses must be capable of
confirmation or disconfirmation through empirical observation of language in use.”
(Leech, 2000:685). Linguistic theories describe and characterise the rules and
classifications within the language. Data-based or usage-based models can then
utilise these descriptions to demonstrate “….how it is deployed in [….] the
language”, and indeed whether it is deployed in the language (Leech, 2000:691). A
particular construction may well be entirely permissible, yet rarely used, and the
corpus-based analysis could help reveal such information.

The role of frequency is conspicuously important in terms of a quantitative
study, but this does not detract from the usefulness of corpora in other respects. In
addition, the corpus approach can contribute valuable qualitative information by
means of contextual data. For example, some words may lose certain senses over
time (semantic narrowing), mono-semantic words may gain additional senses
(semantic broadening), and some of these senses may become more or less popular
with time. In this way, corpus data can be used both quantitatively and qualitatively
to provide empirical support for a given theory or hypothesis. Corpora typically use
quantitative data from vast amounts of authentic language, and produce statistics at a
macro-level, which enables language trends to be identified. However, corpora also
allow micro-level analysis at the sentence, word or even affix level, and thus make
corpora an ideal tool for comprehensive and all-inclusive investigation of
morphological productivity.
3.1.1 Why the British National Corpus?

The British National Corpus (BNC) is a body of over one hundred million words. The source of the words is a collection of over four thousand texts of British English, which are stored in electronic format, and include examples of both written and spoken English. The texts are taken from a variety of different domains and from a variety of different text medium in order “to reflect the widest possible variety of users and uses of the language” (BNC Handbook, 1998: Preface). As Leech et al state:

The BNC is a finite, balanced, sampled corpus. It is possible to extrapolate from corpus frequencies to inferences about the language as a whole, because the compilers have taken great pains to sample different kinds of speech and writing (e.g. conversation, novels, news reporting) broadly in accordance with their representation in everyday language use.

(2001:1)

Likewise, as Kennedy observes, the BNC “was designed to be representative of British English as a whole and not just one particular genre, subject field or register.” (1998:50). Thus, at the time of the inception of this research, the BNC appeared as the most comprehensive, representative, and large-scale corpus, which was readily-available, ‘fit for purpose’ and usable.

3.2 Which negative prefixes and why?

It has been established that a corpus-based approach is most appropriate for a study of morphological productivity in contemporary English, and that the BNC provides the most suitable corpus to use, this section now turns to the question of which negative prefixes make the best candidates for investigation.

Affixal or morphological negation is realised almost exclusively in English by the use of negative prefixes. The one possible exception to this is the suffix -less,
which forms adjectives from nouns and verbs, and carries with it the general meanings ‘without or lacking’ (e.g. hopeless, speechless, childless), ‘free from’ (e.g. painless, careless), and ‘not able to do, or not able to be done’ (e.g. countless, tireless). However, with its inherent adjective-forming characteristics, the suffix -less also serves a grammatical function, and changes the class of the base word, whereas the addition of a negative prefix effects only a semantic transformation on the base; the base word is modified by the addition of a prefix, but the word class is maintained (Funk, 1971:365). On this basis, since -less exhibits a quite different character to that of other forms of affixal negation, it is disregarded for the purposes of this study, which, as a result, focuses solely on prefixal negation.

It is, therefore, necessary to determine which of the many prefixes of English bring about negation, and which of these are of most interest to the present study. On this basis, a list of potential negating prefixes was taken from the Collins Cobuild English Guides 2: Word Formation (1991), since this provides an extremely comprehensive list. The list, therefore, comprises the prefixes: a- (an-\(^1\)), anti-, contra-, counter-, de-, dis-, dys-, ex-, extra-, in- (il-, im-, ir-\(^2\)), mal-, mis-, non-, para-, sub- and un-. Secondly, though not included in the Collins Cobuild English Guides, the prefix e- was also added to this list as a potential negative prefix for this study. Whilst not an especially well-known prefix, intuitively it seemed worthy of inclusion, at least at this initial stage, since it appears to denote negation. Finally, as this study considers prefixal negation in the restricted sense of the concept, the list was then edited according to the principles set out below.

Firstly, the prefixes anti-, contra-, counter- and de- were disregarded, since these four intrinsically carry semantic characteristics different to that of simple

---

\(^1\) The a-prefix appears as an- before a subsequent vowel, for example: anaerobic.

\(^2\) The prefixes il-, im- and ir- are all variants of the prefix in-: il- occurring before ‘l’ (illogical), im- occurring before ‘b’, ‘m’ and ‘p’ (imperfect) and ir- occurring before ‘r’ (irregular).

For similar reasons, dys-, mal- and mis- were discounted. Dys- connotes implications of ‘abnormality, impairment’: dysfunction; or ‘difficulty, painfulness’: dysmenorrhoea (New Penguin English Dictionary, 2000), in a similar fashion to mal- (malpractice) and mis- (misjudge), which have purely pejorative functions (Mazzon, 2004:112). Though all three prefixes could be said to possess a degree of negation, their sense is more in terms of expressing an evaluative judgement.

The prefixes extra- and para-, in words such as extraterrestrial and paranormal might be considered as simply meaning ‘not’ in the sense of ‘not terrestrial’ and ‘not normal’, but the sense of the prefixes here indicate ‘beyond or outside of’. Similarly, subnormal may be considered as meaning ‘not normal’, though the sense of sub- in fact denotes ‘below, beneath, under or incomplete’. Whilst the prefix e- was initially included, it is now excluded, along with the prefix ex-, since, on closer inspection, the two also signify the deprivation of a state or quality, and to a lesser degree ‘lacking or without’, though in the case of the prefix

---

3 “extra-terrestrial = existing or originating outside the earth or its atmosphere; hence as n., a being living or originating beyond the earth” (OED) (my italicising). Likewise, there is also the existence of a contrasting form, unterrestrial, which uses the prefix un- in a “purely negative” sense (OED – entry for un- Sense 7a).
e- (such as in *ecaudate, edentate*), there is an additional sense, which contrasts with the simple negative meaning, of ‘out or away’. In the case of the prefix *ex-*, though it may possess some sense of a degree of deprivation, for example in *excommunicate*, as a synchronically productive prefix, there is little to imply the negation of a positive term with its interpretation as ‘former’ (for example, as in *ex-President*). For a complete list of the potential negative prefixes from which the final selection was made, see Table 3.1 below.

<table>
<thead>
<tr>
<th>Negative Prefix</th>
<th>Example</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>a- (an-)</em></td>
<td>atypical (anechoic)</td>
<td>not / negation / without / opposite to</td>
</tr>
<tr>
<td><em>anti-</em></td>
<td>anticlerical, antifreeze</td>
<td>opposing / against counteracting</td>
</tr>
<tr>
<td><em>contra-</em></td>
<td>contradistinction, contraflow</td>
<td>contrary / contrast</td>
</tr>
<tr>
<td><em>counter-</em></td>
<td>counterproductive, counter-strike, counterfoil</td>
<td>opposing / retaliatory / corresponding</td>
</tr>
<tr>
<td><em>de-</em></td>
<td>depopulate, delouse, devalue</td>
<td>opposite / removal / reduction</td>
</tr>
<tr>
<td><em>dis-</em></td>
<td>dishonest, disappear, dislike</td>
<td>not - negation / reversal</td>
</tr>
<tr>
<td><em>dys-</em></td>
<td>dysfunction, dysmenorrhoea</td>
<td>abnormality / impairment / difficulty / painfulness</td>
</tr>
<tr>
<td><em>e-</em></td>
<td>ecaudate, edentate</td>
<td>deprivation / out / away</td>
</tr>
<tr>
<td><em>ex-</em></td>
<td>excommunicate, ex-President</td>
<td>deprivation / former</td>
</tr>
<tr>
<td><em>extra-</em></td>
<td>extraterrestrial</td>
<td>beyond / outside</td>
</tr>
<tr>
<td><em>in-</em> (il-, im-, ir-*)</td>
<td>incomplete (illegal, impossible, irregular)</td>
<td>not / negation</td>
</tr>
<tr>
<td><em>mal-</em></td>
<td>malpractice</td>
<td>bad / badly / faulty / deficient</td>
</tr>
<tr>
<td><em>mis-</em></td>
<td>misjudge</td>
<td>badly / wrongly / unfavourably</td>
</tr>
<tr>
<td><em>non-</em></td>
<td>non-British</td>
<td>not / negation</td>
</tr>
<tr>
<td><em>para-</em></td>
<td>paranormal</td>
<td>beyond / outside</td>
</tr>
<tr>
<td><em>sub-</em></td>
<td>subnormal</td>
<td>below / beneath / under / incomplete</td>
</tr>
<tr>
<td><em>un-</em></td>
<td>unhappy, untangle, untidy</td>
<td>not / negation / reversal / opposite of</td>
</tr>
</tbody>
</table>

As the table shows, most of the prefixes have a sense which is discrete from that of negation, whereas only five prefixes strictly reflect negation: the prefixes *a-*, *dis-*,
in- (along with its variants *il-*-, *im-*-, *ir-*), *non-* and *un-*.

It should, however, be noted that verbal *dis-* examples (e.g. *disappear*, *discredit* and *dislike*) are included in the database, despite the fact that it could be argued that some examples of verbal *dis-*

whilst they may demonstrate negation, also go beyond negation. For instance, some examples of verbal *dis-* also have a supplementary sense to that of negation. *Disappear*, for example, negates *appear* and creates an opposition, but also included in this negation is a supplementary reversion sense. The rationale for this inclusion, however, is that all senses represent clear cases of antonymy or oppositeness. Cruse, in his comprehensive categorisation of antonyms, includes reversionals under the category of negation in terms of their opposition (1986:226), and indeed, a pair of morphological antonyms corresponds to a positive term and its negated counterpart. For example, “when asked for the opposite of *obey*, the invariable response is *disobey*, presumably because *obey:* *disobey* are ‘better’ opposites” than other potential antonyms (Cruse, 1986:201). Likewise, it should also be acknowledged that, to some extent, *un-* could similarly be treated as having two different senses. For example, the sense of *un-* in *untangle*, along with other verbal examples such as *unblock*, *uncork*, *unload*, etc., also have a reversional sense, which supplements the sense of pure negation. However, this reversional sense is not strictly a discrete meaning, rather, as with *dis-* , the use of the prefix *un-* in verbal examples is a sense which is encompassed by the negative sense. For this reason, *dis-* and *un-* prefixed verbs are included in this study.

Several linguists (Funk, 1971:368; Zimmer, 1964:26; Bauer, 1983:218) discount the inclusion of the negative prefix *a-* in terms of its synchronic relevance. As Funk claims, “There are not many words with this prefix [*a-*] in English, and even fewer remain after selection on a strictly synchronic basis”. He goes on to add
that “….they are, with very few exceptions, scientific terms….” (1971:368). Despite this assertion, many, if not most, of Funk’s examples of negatively a-prefixed words are not in fact from the realms of science (acanonical, acritical, amechanical, apopular, etc.), and even if they were, this should not preclude them from a synchronic category of negating prefixes. As is demonstrated in Hulse (2002), many of the words negatively prefixed with a- are not specifically from the field of science, for example, acerebral, amoral, asocial, atypical, etc. Furthermore, the negating a- prefix “remains sufficiently fertile to generate new formations” (Hulse, 2002:118). Indeed, other researchers, such as Matthews (1991), Crystal (1995), Kwon (1997), Plag (2003) and Mazzon (2004) also believe that the negating a- prefix is relevant in an investigation of affixal negation, albeit that it often occurs in academic or scientific settings. On this basis, the negating prefixes to be examined in this thesis are: a-, dis-, in- (its variants il-, im-, ir-), non- and un-.

3.3 Creating a database of negatively-prefixed words

Having identified the negative prefixes to be utilised in this study, the next stage of the research was to extract all the examples of negatively-prefixed words in the corpus in order to create a definitive database. The BNC tags all the words for word class and lists all the types, along with their token frequencies. Likewise, with the database saved as Excel files, there is also the facility for the data to be stored alphabetically, according to frequency or according to word class. The next step, therefore, was to use the BNC as the source to extract all words which begin with the negative prefixes a- (and an-), dis-, in- (il-, im- and ir-), non- and un- in order to compile a list for each prefix, and to show the frequency of each word in the corpus.

4 Acerebral is not, in fact, a medical or scientific term, but means ‘unintelligent or unthinking’.
and how the *BNC* tags it for word class (adjective, adverb, noun, preposition, verb). These particular word classes were chosen, since they are ‘open’ word classes and, as such, allow for the creation of new formations which incorporate productive affixes.

A number of earlier studies, such as Funk (1971) and Baldi et al (1985), limited their research to negative adjectives. However, it was felt that negatively-prefixed words from other word classes are significantly represented in the language, and subsequently in the database, and therefore this research should not be restricted solely to adjectives. Perfectly familiar examples for each of the prefixes from other word classes might include: *asymmetrically* (adv), *asexuality* (noun), *discontinuously* (adv), *dissatisfaction* (noun), *disappear* (verb), *illogically* (adv), *immaturity* (noun), *inactivate* (verb), *non-verbally* (adv), *non-co(-)operation* (noun), *unaffectionately* (adv), *unimportance* (noun), and *unbalance* (verb).

### 3.3.1 Use of software to identify negatively-prefixed words

Adam Kilgarriff provided suitable software: Sketch Engine, a tool which automatically identifies the candidate words from the *BNC* for inclusion in the database ([http://www.sketchengine.co.uk](http://www.sketchengine.co.uk)). Sketch Engine has the advantage of being able to group negatively-prefixed words together for each of the prefixes, according to their word class and frequencies in the *BNC*. Then, stored as Excel files, each list can be sorted alphabetically, and within that alphabetical list, sorted according to word class. The lists can also be ordered according to word frequency. The above facilities are not features available via standard *BNC* software, though the *BNC* can be used to view the specific context of any sample word in the database. Thus, Kilgarriff’s shortcut software was used to create lists of words negatively-prefixed with *dis-*-, *in-*-, *non-* and *un-*.
The Sketch Engine software was not, however, used to create the a-prefix list because of the problem of accurately being able to identify a-prefix words. Words prefixed with a- tend not to be receptive to automatic identification since, as a single letter, software simply brings up all words which begin with the letter ‘a’. For the creation of the a-list, therefore, the BNC ‘Word Query’ function was used, which brings up all words (approximately 31,000) that begin with ‘a’, along with all their frequencies. It does not, however, show the word classes of the words on the list. Despite extracting such a large number of words beginning with the letter ‘a’, large numbers of these words (i.e. non-prefix words), were relatively-easy to delete manually. Therefore, for example, able, accident, after, apply, etc. were deleted, along with all words derived from these, such as: ability, able-bodied, ableism; accidental, accidentally, accident-prone; afterlife, aftermath, afternoon, etc. Words which could be categorised as prefixed but were not negatively-prefixsed, such as ablaze, adrift, aloud, asleep, etc., were also manually deleted. As a result, the a-list was reduced to a more manageably-sized list of a few hundred words, and the word classes were then added to the list manually. These criteria and several others, to be described below, were also applied to the other four negatively-prefixsed word lists for dis-, in-, non- and un-.

3.3.2 Editing the database

A quantitative approach, as is taken here, enables the compilation of large amounts of data, would have been impracticable without the use of electronic corpora. Nevertheless, the word lists required editing to eliminate and/or merge entries in a more discerning way than the corpus or the query system would allow. In other words, a qualitative element would be introduced to support, and work in
conjunctio
cn
with, the quantitative measure, whereby the automatically-generated lists would be manually edited.

The word lists produced required manual editing for a variety of reasons. As with the _a_- list, the _dis-_, _in-_, _non- _and _un-_ lists included all words beginning with that combination of characters so, for example, _disaster_, _internal_, _none_ and _university_ occurred in the lists but were obviously not negatively-prefixed words. The lists were manually searched for all such examples, along with all their derivatives, and these were deleted. The word lists were also manually searched for obvious spelling or typographical errors, for example, _disbeleiver_ and _unfortunatly_, and these were merged, along with their frequencies, with the correctly-spelt versions of the words. Likewise, accepted spelling variations were merged, along with British and American English versions of the same word. These might include, for instance, _disorganise_ and _disorganize_, which would be edited and merged as _disorgani(s/z)e_, and _discolour_ and _discolor_, which would appear as _discolo(u)r_. Hyphenated and non-hyphenated forms of words were also merged in examples such as, _non-periodic_ and _nonperiodic_, which were listed as _non(-)periodic_; and _un-numbered_ and _unnumbered_, were merged and included as _un(-)numbered_. Hyphenated compound words were eliminated, but their frequencies (often 1) were added to the simple negatively-prefixed word. For example, _implausible-sounding_ (Freq. 1) was deleted, but its frequency was added to the entry _implausible_ (Freq. 157) which then gives a frequency of 158 for _implausible_. Similarly, in cases where a word was listed as a separate word simply because it began with a capital letter (for example, where the word’s context showed it at the start of a sentence), this example was also merged with those that began with a lower case letter.
The negatively-prefixed word lists were also inspected for incorrect word class tagging and, if doubtful, were checked for context in the BNC. The item was then corrected if found to be incorrectly tagged but, again, its frequency was added to the correctly-tagged version of the word. Likewise, where the meaning of a word was not known or may have been considered ambiguous, the context was checked in the BNC, and the meaning was verified via the OED. For example, disputative was listed as an adjective with a frequency of 3. In isolation, however, an ambiguity exists, since disputative could be a negative form of putative, or it could be ‘something which is capable of being disputed’, in which case, it is not a negatively-prefixed word. The context was checked with the BNC, it was found to have the latter meaning, ‘something capable of being disputed’ and, therefore, it was deleted from the word list, since it was a derivative of dispute and not a negative form.

Ambiguities were also discovered which, on first inspection, did not appear to be ambiguities at all. One example of this was the noun disembowelment. Initially, it might be classed as the negative form of embowelment which, according to the OED, does exist, but in fact embowelment means ‘the action of disembowelling’. On this basis, therefore, disembowelment was deleted from the list, since it would not be deemed a negative form of embowelment. Similarly, examples such as discombobulate and discomknockerate / discumknockerate were also deleted from the lists since, though they appear to have recognisable components (negating dis-, verb-forming -ate), they do not in fact have positive forms (*combobulate, *comknockerate / *cumknockerate). Finally, where a prefix has been applied to a word but is not in fact strictly a negative prefix, these were also deleted. For example, the prefix in the word dissever has the sense of ‘separation’ rather than negation, and the prefix is described as being used “with verbs having already a
sense of division, solution, separation, or undoing”, whereby “the addition of dis- was naturally intensive” (*OED Online*). Therefore, before retaining a prefixed word in the lists, the actual meaning of the prefix in that specific example was ascertained via the *BNC* contexts and *OED* entries.

For consistency of editing, therefore, a nine-point ‘Inclusion and Exclusion Criteria’ list was prepared, and was strictly applied to all the lists of negatively-prefixed words. A manual, qualitative component was, therefore, employed with regard to the editing of the data. The Inclusion and Exclusion criteria were as follows:

1. Deletion of accidental character combination i.e. non-negatively prefixed words (e.g. *disaster, internal, none, university*, etc.).

2. Merging of spelling errors (e.g. *disbeleiver, unfortunately*) and spelling variations (e.g. *disorgani(s/z)e, discolo(u)r*, etc.).

3. Merging of hyphenated and non-hyphenated words (e.g. *non(periodic) and nonperiodic, un-numbered and unnumbered*).

4. Merging of capitalised and uncapitalised words (which are listed separately).

5. Deletion of hyphenated compound words with frequency correction for first component (e.g. *implausible-sounding* (Freq. 1) deleted; *implausible* (Freq. 157) amended to *implausible* (Freq. 158).

6. Deletion of semantically-ambiguous words in which the prefix does not serve a negating function though appears to be (e.g. *dissever, inflammable, disembowelment*, etc.).

7. Ambiguous examples (e.g. *disputative – negative form of putative* or ‘capable of being disputed’)?) were dealt with manually via examination of context.

---

3 When editing the *a*-prefix list, examples such as *acephalous, amnesia, and anorexia* were deleted. Despite the ultimate Greek etymology of the words, the negating prefix: *a-, an-* (‘lacking’) is not negating a current English word, since *cephalous* (‘head’), *mnesia* (‘memory’), and *orexia* (‘appetite’) do not constitute independent words in English.
8. Deletion of nonce words which have the characteristics and effect of a negative word form, though no positive form exists (e.g. discombobulate, disc(o/u)mknockerate, etc.).

9. All words found to be incorrectly tagged were correctly re-tagged.

3.3.3 **Accuracy of the BNC and Sketch Engine word-class tagging**

Kilgarriff’s software generally identifies word classes relatively accurately, since during the course of the editing, thousands of examples were examined overall in terms of their contexts, and the accuracy of the word class tagging was able to be checked and confirmed. It was ultimately decided, therefore, to generally rely on the word-class tagging of Sketch Engine, rather than re-tag the whole database manually.

3.3.4 **The final database**

Finally, when the *a*, *dis*, *in*, *non* and *un* lists were fully edited and the inclusion and exclusion criteria applied, the word classes were manually added to the *a*-prefix list in order that the format of all the lists of prefixes were identical. This was achieved by means of checking the contexts of the negatively-prefixed *a*-words where the word class was not obvious or there may have been some ambiguity. Thus, the database was created, and the lists could be sorted alphabetically, according to word class and according to frequency. The total number of individual negatively-prefixed words in the database (the types) amounted to 9,121, which could be broken down into 150 negatively-prefixed *a*-words, 429 words prefixed with *dis*-, 922 words prefixed with *in*-, and 3,784 and 3,836 words prefixed with *non*- and
un- respectively. (See disc at end of thesis – The Database). The total number of negatively-prefixed types is illustrated in Table 3.2 below, and closer analysis of this table and the database as a whole takes place in the analysis chapter to follow.

<table>
<thead>
<tr>
<th>Total No. of negatively-prefixed words</th>
<th>Adverbs</th>
<th>Adjectives</th>
<th>Nouns</th>
<th>Prep’s</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>150</td>
<td>16</td>
<td>88</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>dis-</td>
<td>429</td>
<td>27</td>
<td>143</td>
<td>166</td>
<td>0</td>
</tr>
<tr>
<td>in-</td>
<td>922</td>
<td>193</td>
<td>351</td>
<td>358</td>
<td>0</td>
</tr>
<tr>
<td>non-</td>
<td>3,784</td>
<td>70</td>
<td>2,525</td>
<td>1,188</td>
<td>0</td>
</tr>
<tr>
<td>un-</td>
<td>3,836</td>
<td>338</td>
<td>2,879</td>
<td>404</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>9,121</td>
<td>644</td>
<td>5,986</td>
<td>2,162</td>
<td>1</td>
</tr>
</tbody>
</table>

In terms of the frequencies of all the individual negatively-prefixed types in the database (the tokens), this amounts to 343,956 tokens in total. These figures are reflected in Table 3.3 below, which also shows the distribution of the frequencies between the different word classes for each of the prefixes.

<table>
<thead>
<tr>
<th>Total frequencies of negatively-prefixed words</th>
<th>Adverbs</th>
<th>Adjectives</th>
<th>Nouns</th>
<th>Prep’s</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>3,347</td>
<td>70</td>
<td>1,817</td>
<td>1,460</td>
<td>0</td>
</tr>
<tr>
<td>dis-</td>
<td>38,368</td>
<td>661</td>
<td>6,225</td>
<td>18,241</td>
<td>0</td>
</tr>
<tr>
<td>in-</td>
<td>108,791</td>
<td>15,580</td>
<td>70,079</td>
<td>22,372</td>
<td>0</td>
</tr>
<tr>
<td>non-</td>
<td>22,466</td>
<td>292</td>
<td>16,287</td>
<td>5,884</td>
<td>0</td>
</tr>
<tr>
<td>un-</td>
<td>170,984</td>
<td>18,340</td>
<td>125,210</td>
<td>14,231</td>
<td>3,902</td>
</tr>
<tr>
<td>Totals</td>
<td>343,956</td>
<td>34,943</td>
<td>219,618</td>
<td>62,188</td>
<td>3,902</td>
</tr>
</tbody>
</table>

6 It should be noted that the 3,836 un- types are divided between 338 adverbs, 2,879 adjectives, 404 nouns, 214 verbs and also 1 type which is a preposition (unlike). This study, however, focuses on the open word classes and the single prepositional type, unlike, is not included in the analysis.
Again, this table is examined in greater detail in the analysis chapter, with both tables (3.2 and 3.3) forming the basis for the analysis of the type and token measures of productivity.

Before analysing the database in closer detail, this thesis now turns its attention to the historical and semantic factors that have been shown to influence the use of these five negating prefixes. As outlined in Chapter 2, these factors can affect the different nuances in meaning of the five prefixes. Firstly, some relevant historical factors are examined.

3.4 Historical factors affecting the negative prefixes

It is the belief of Baldi et al (1985) that “(n)egative prefixing is a fertile testing ground for theories of morphological productivity”, and that etymology, for example, plays a part in the choice of one prefix over another (33). They claim that “native speakers have access to and make use of etymological information about roots and prefixes before assigning prefixes to roots” (34). In their study which involves psycholinguistic experiments, subjects were asked to identify negatively- prefixed adjectives as either native English or foreign words. Their findings suggest that the subjects classified the words’ origins “with a high degree of accuracy” (Baldi et al, 1985:54). Thus, they concluded, the history of a specific negating prefix has some bearing on its current productivity, depending on the time and the circumstances of its introduction into the language.

The negating a- prefix came into English from Greek and was, until comparatively recently⁷, only attached to words of Greek origin, for example

---

⁷ According to the OED Online, the first recorded negation produced in English using the negating a- prefix was during the Renaissance in 1646, when the ‘negative’ noun adiabolist was formed from the positive term diabolism. It was not, however, until 1835 that the prefix was added to a word of
The prefix is, however, now also used to negate words of Latin origin, for example, *amoral, asexual, asocial*, etc. The prefix was originally introduced into English already incorporated into Greek loan words, but it was not until the start of the 17th Century that it was first used as a true negating prefix in English (OED Online). Its popularity as an English prefix in its own right came at the time of the Renaissance, a period which was characterised by the exploration of the world and the individual (Hulse, 2002:24). There was, at this time, a concentration on the active rather than passive lifestyle, in addition to an overriding change in tenor from religious to secular thought. Initially, the *a-* prefix was used mainly in religious, scientific and medical terminology, as the Renaissance was also a time of scientific and engineering advancement, along with the development of new artistic and literary styles. As a result of these widespread developments, many new nominal and descriptive terms were needed and, in time, this need for new terminology was also reflected in the need to create antonymous or negative terms to complement the positive ones. Indeed, the number of nouns and adjectives in the database which are prefixed with *a-* could again be ascribed to how the negative *a-* prefix was first recorded in English, and how it has been utilised since. As Marchand points out, “The English [*a-* prefix] coinages are all derivatives from a substantival basis [derived from nouns], though in practice they are often analysed as opposites of unprefixed adjectives.” (1969:140). A number of linguists, such as Zimmer (1964), Funk (1971), Bauer (1983) and Kwon (1997) dismiss the negative *a-* prefix as synchronically irrelevant. Funk proposes that there are very few words in English with this prefix (1971:368). Similarly, Kwon states that “(w)ords with this prefix are of very limited occurrence in English” (1997:16).

Latin origin, *normal* to produce *anormal*, though this was subsequently altered to the now more familiar *abnormal* (OED Online, 2001).
The negative prefixes *dis-* and *in-* have somewhat similar profiles, both diachronically and synchronically, and their status as prefixes from an earlier stage of the language may account for certain aspects of their distributional similarities. Both *dis-* and *in-* are of Latin origin, and according to the *OED*, *in-* was first introduced into the language circa 1225, with its first recorded occurrence as the noun *ingratitude*, and has continued to be attached only to words of Latin or French origin. In a privative sense, *dis-* was first recorded in English circa 1300, again as a noun: *dishonour*. Although Marchand (1969) describes *dis-* as a prefix which combines with words that “have a learned or academic tinge” (161), both *dis-* and *in-* have a tendency to combine with more common words of English; words which are perhaps more likely to be more frequently used in everyday language, such as *(dis)able, (dis)agree, (dis)appear, (dis)honest, (il)legal, (im)possible, (in)dependent, (ir)relevant*, etc. In many cases, however, the base word’s choice of negating element may appear arbitrary:

The application of affixes is totally dependent on lexical diffusion, so each prefix had to ‘conquer its territory’, so to speak. It is thus that *impossible* won over *unpossible*, *uncertain* over *incertain*, *unjust* over *injust*, and so on for dozens of other cases. The factors that rule such choices may well range from euphony and articulatory motives to cultural prestige of some trend-setting writers or early dictionary compilers, and in this sense, as mentioned, each word is a case apart. (Mazzon, 2004:74)

*Un-* is the only native negating prefix of English and, although it is first recorded as early as the 8th Century, many examples were subsequently lost with the disappearance of Old English in the 12th Century. Since the 14th Century, however, “the productivity of the prefix *un-* has been immense” (Kwon, 1997:20).
Historically, *un-* has always been associated with adjectival negation, and its ability to combine with adjectives of both native and foreign origin.

At the time of the re-introduction of the prefix *un-* to the language around the 14th Century, the *in-* prefix was only attached to words derived from Latin. However, since *un-* was an indigenous prefix and could be freely attached to words to create the negative form, the *in-* prefix began to lose out to *un-*.

Likewise, in more recent years, the prefix *un-* has spread to words which were formerly only prefixed with *in-*, particularly in cases where the prefix *in-* has developed a more specific sense, as opposed to a purely negative sense. For instance, in the case of *inhuman* and *unhuman* respectively, *inhuman* has developed a more common meaning of ‘lacking humane feelings; cruel or brutal’, whereas *unhuman*, while this can also mean ‘inhumane’, can simply mean ‘not human or not pertaining to humans’ (*OED Online*). The more common and everyday nature of words which incorporate *dis-* and *in-*, along with their extremely long history in the language, may well have had an influence on their current standing as negative prefixes. Although *dis-* and *in-* would once have been extensively used for negation, since the introduction of other prefixes into the language, *dis-* and *in-* have continued to lose out over time to the more contemporary negative prefixes.

The prefix *non-* was introduced into English from Latin around the 14th Century but was initially used almost exclusively for legal terminology, very often in nominal constructions (Marchand, 1969:179; Kwon, 1997:19; Mazzon, 2004:111). By the end of the 16th Century, new *non-* words tended to come from the fields of “philosophy, religion and political history”, and its use was subsequently extended to include adjectival constructions and can now be freely attached to adjectival bases (Marchand, 1969:179). Apart from its historical standing, the neutral status of
non- as a negating element may well explain its high number of nominal and adjectival types (see also Section 3.5 which relates to semantic factors). *Non-* was rarely found in literature in the 14th, 15th and 16th Centuries, but by the 17th Century it began to be used in scientific vocabulary. Marchand states that, although *non-* was previously used in verbal combinations between the 15th and 17th Centuries, these derivations have since become obsolete (1969:180). Lieber confirms this view when she states “(n)on- does not attach to verbs.” (2007:114). The current database of this research, however, does demonstrate one verbal type in *non-*; again a legal term, in the form of *nonsuit*. *Non-* began to be more freely used in scientific terminology, and has gained more freedom in terms of the bases with which it could combine (Algeo, 1971:88-90; Funk, 1971:372; Marchand, 1969:180; Bauer, 1983:279).

This concludes a brief history of the five prefixes. The purpose of this section has been to chart the paths taken by these prefixes en route to their current status as familiar and everyday components of the English Language. The next section reports a small historical study to examine whether the introduction of a prefix into the language might ultimately have an effect on its current status in the language.

### 3.4.1 Small historical ‘snap-shot’ study

It could be argued that the date of introduction of a prefix into the language may have an influence on the productivity of that prefix. In order to test this notion, a small ‘snap-shot’ study was undertaken, whereby the highest-frequency word for each of the five prefixes was taken from the database, as a sample and as a typical and common example of each prefix. The date chart in the *OED* was used to establish a date of introduction for each negative word and its positive counterpart. The words examined, therefore, were: *aesthetic* (1832) and *anaesthetic* (1846);
appear (1375) and disappear (1530); dependent (1420) and independent (1611); existent (1561) and non-existent (1656); and employment (1598) and unemployment (1888). The sample words demonstrate that those which are subsequently negatively-prefixed with dis- and in- (appear (1375) and dependent (1420)) are introduced into the language at an earlier period than those which will subsequently be negatively-prefixed with a-, non- and un- (aesthetic (1832), existent (1561) and employment (1598)). The negative counterparts disappear and independent, are first introduced around 1530 and 1611 respectively, whereas non-existent (1656), anaesthetic (1846) and unemployment (1888) appear at later stages of the language. Figure 3.1 shows, therefore, the date of introduction of the positive forms of the words, along with the subsequent introduction of their negative counterparts.

Figure 3.1. Introduction of the positive and the negated forms of the highest-frequency negatively-prefixed words from the database

Once dis- and in- were used to negate words, this would perhaps block the other negative prefixes being used for those same words, unless another sense of the
word was required or introduced. In earlier times, *dis*- and *in*- would have been attached to relatively common everyday words. Indeed, these are words which are typically still common words of the language. Thus, it is proposed that the older prefixes may be more likely to be shown as the least currently productive. Indeed, it may well prove to be the case that historical factors may have an important effect on the current choice of one negative prefix over another.

### 3.5 Semantic factors affecting the negative prefixes

Since each prefix has its own idiosyncratic characteristics and nuances, this section reviews the semantic features which may have an influence on the negative prefixes in terms of their use and productivity. For example, certain prefixes have a tendency towards contradictory negation, whereby the type of negation is binary and non-gradable (e.g. *non-professional*, *unmarried*). Other prefixes tend towards a contrary type of opposition, or opposition which may be gradable (e.g. *impolite*, *unhappy*). Different linguists describe the negative prefixes differently in terms of their semantics. Thus, the ideas of Marchand (1969), Funk (1971), Kwon (1997), Mazzon (2004), Lieber (2007) and Hamawand (2009) are described with regard to the semantics of each of the negative prefixes.

#### 3.5.1 *A*-prefix semantics

Marchand (1969) defines the meaning of the *a*- prefix as ‘without’, ‘devoid of’ or ‘not affected or characterised by’ or ‘not’, and describes it as an adjectival prefix, and that all its derivatives are “from a substantival basis, though in practice they are often analysed as opposites of unprefixed adjectives.” (140). Marchand provides examples such as *achromatic* and *aseptic*, and describes the type as contradictory
opposites. Funk (1971) also describes the *a-* prefix as contradictory in type, with a few examples which are privatives (e.g. *acaudal* ‘having no tail’) but suggests the prefix is somewhat irrelevant synchronically and that there are very few formations in English which incorporate the prefix. As was stated earlier, Funk is of the opinion that virtually all formations are science-related terms, though most of his examples are not in fact scientific (368-369). Kwon (1997) argues that the *a-* prefix relates to highly specialised technical language (182), and also proposes that in cases where a base word may take a variety of prefixes, the *a-* derivation tends to refer to people, as opposed to behaviour or actions. For example, *amoral* usually refers to a person, whereas *immoral* has a tendency to relate to a person’s behaviour (178). Mazzon (2004) claims that the semantics of the *a-* prefix refers to a ‘lack of’ something and that it occurs in “learned vocabulary” (111).

Whilst examination of the database certainly shows up scientific and technical *a-* prefixed terminology, many of these concern natural science and medicine. Nevertheless, a considerable number of these examples are also used in more commonplace, non-scientific contexts. Such examples include *amoral, asexual, asocial, asymmetrical, asynchronous* and *atypical*. Examples like *asynchronous* may in fact be gaining more widespread and general popularity in respect of increased Internet usage in terms of ‘synchronous’ and ‘asynchronous’ types of online courses and training (i.e. real-time and non-real time events). Similarly, and though still used in scientific and medical contexts, most of us have encountered the terms *asymmetric* and *asymmetrical* in maths lessons at school. Indeed, the database shows *asynchronous, asymmetric* and *asymmetrical* amongst the highest frequencies for *a-* prefixed words with frequencies of 140, 139 and 114, respectively.
3.5.2 *Dis-* prefix semantics

As cited in Kwon (1997), the *OED* states that the prefix *dis-* now has three distinct, though ultimately related, senses. The first of these is a straightforward negative sense meaning ‘not’; the second is a “privative or reversative sense” meaning ‘undoing or reversing an action’; and the third sense, the original “etymological sense” of ‘apart or in different directions’, which is outside the scope of this current research (17). Marchand proposes that most current *dis-* derivatives are contradictory oppositions (the first sense), though there are some contrary opposites and reversatives, and similarly, also proposes that there are a number of *dis-* verbs which combine with -en- and -em- (*disembody, disentangle, disentwine*) which are still current, though these are not necessarily frequent examples (1969:160-161). The database, in fact, lists twelve of these *disem-*/*disen-* types, which almost all have minimal frequencies. Only *disenchant* (frequency 26), *disenfranchise* (frequency 54) and, as mentioned, *disentangle* (frequency 170) have a notable number of occurrences. Other than these verbal types, Marchand states that *dis-* words are common with nouns and adjectives (1969:161).

Funk (1971) also concurs with Marchand (1969) that *dis-* types are either contrary (*dispassionate*) or contradictory (*disloyal*), and points out that certain types, such as *disinterested*, are semantically restricted (cf. *uninterested*) (369), a view which is also supported by Kwon (1997:184-185). Both Funk (1971) and Kwon (1997) admit that a semantic distinction between *disinterested* and *uninterested* existed in the past, though this distinction is gradually being lost (Funk, 1971:369; Kwon, 1997:184-185). Kwon examines the pair in greater detail, and determines that semantically *disinterested* has the sense of impartiality and is most commonly used attributively. *Uninterested*, on the other hand, implies ‘not interested’ and is more
often used predicatively (1997:185). The view is also that *disinterested* is losing out to *uninterested*, since *un-* words are more associated with “words of general currency” (Marchand, 1969:161). Indeed, Marchand claims that *disinterested* has already been lost to American English (1969:161). However, according to the BNC, *disinterested* has a frequency of 172, whereas *uninterested* only has a frequency of 138, therefore the claims of Marchand (1969), Funk (1971) and Kwon (1997) are not necessarily supported by the current research.

Both Funk and Lieber, almost forty years apart, propose that the choice of prefix is ultimately determined by the semantic features of the base word, rather than those of the prefix itself (Funk, 1971:382; Lieber, 2007:115). Kwon (1997) further proposes that *dis-* also has the meaning ‘less than’ or ‘lacking’ in an impartial, objective sense, for example, when *disinterested* is used attributively (185). Mazzon claims that, in addition to its frequent use in nouns and verbs, *dis-* also combines to form adverbs (2004:111). Funk also goes on to propose that with *dis-* “(t)he lack of syntactic derivatives of the prefix in present-day English corresponds to the fact that its productivity certainly equals zero today.” (1971:369).

### 3.5.3 *In-* prefix semantics

The negative prefix *in-* is described as one which forms either contrary or contradictory opposition (Funk, 1971:370; Kwon, 1997:187-192; Lieber, 2007:113-115). Some linguists, such as Marchand and Kwon, refer to *in-* almost as a variant of *un-* (Marchand, 1969:170; Kwon, 1997:187), though Marchand proposes that *in-* competes with, but typically loses out to, *un-* (1969:170). Certainly, where no semantic distinction exists, the two forms do appear to act as variants of one another (Kwon, 1997:187). The general consensus is that *in-* almost exclusively combines
with adjectives, but that there are a few nouns that incorporate the prefix (Marchand, 1969:168; Mazzon, 2004:111; Lieber, 2007:113). As shown in tables 3.2 and 3.3, the current database finds otherwise. In terms of types, slightly over one third of all in- types are in fact nouns, though another third is also made up of adjectives. However, with regard to token frequencies, it is certainly adjectives which claim the highest percentage of in- tokens.

It is typically the case that in- prefixed words relate to scientific language and have a tendency to be attached to longer words than those to which the native prefix un- is attached (Marchand, 1969:170). Semantically, where in- and un- compete in the same environment, there is a tendency for in- prefixed words to have more negative connotations, for example in the case of inhuman and unhuman or immoral and unmoral. Inhuman, for instance, means ‘brutal’, whereas unhuman simply means ‘not human’. Immoral means ‘licentious or depraved’, though the meaning of unmoral is simply ‘not moral’ or ‘non-moral’ (Zimmer, 1964:27; Funk, 1971:380-381; Horn, 1989:281; Kwon, 1997:28-29). Likewise, Zimmer (1964), Funk (1971), Horn (1989) and Kwon (1997) all suggest that in- appears to be losing ground in the case of its productivity generally, as well as also seeming to be losing its hold geographically. For instance, although American and British English both still use in- and un-, where there is no semantic distinction, un- prefixed words increasingly become the choice over in- prefixed words in American English. For example, as Kwon claims that, “while irrecoverable is much preferred in British English, unrecoverable is used in both American and British English”, though it is likely that this tendency to choose un- over in- will apply more in British English in time (1997:192).
Kwon claims that, semantically, *in-* prefixed words also have a tendency to develop new senses which, in turn, may have an effect on their productivity (1997:162). A new sense undoubtedly affects how the *in-* prefixed word can be used and also affects the context of any application of the new sense. An example of this could be the adverb *infinitely*. In cases such as this, certain negatively-prefixed *in-* types may come to be more frequent than their ‘positive’ equivalent. For instance, *infinitely* has a frequency of 570 in the database, whereas *finitely* has a frequency of only 14. *Finitely* would typically describe something with the characteristics of ‘limited’, ‘bounded’ or ‘not infinite’. *Infinitely*, on the other hand, is much less frequently used in the sense of ‘limitless’, ‘not bounded’ or ‘not finite’. The use of *infinitely* as an antonym for *finitely* has been overtaken by its use in the sense of an intensifier or hyperbole, whereby in the vast majority of instances in contemporary English, *infinitely* could be replaced with ‘much’ or ‘very’. The contextual examples from the *BNC* demonstrate this more common usage:

(i) He is *infinitely* kind to all things that go into his pictures.

(ii) I also came upon a mysterious animal I had not seen before: much larger than the herring, redder and *infinitely* more expensive.

(iii) It must be said however, that they are *infinitely* more cautious than their predecessors, and on the whole with good reason.

(iv) … it is certainly *infinitely* less serious for Israel's future than the mass popular unrest of the Intifada….

Thus, the development of a new sense may create the effect of increasing the use of the prefix, whereas in reality, the prefix has become an integral part of the whole word and is not truly synchronically analysable into separate components parts. As Jespersen observed almost a hundred years ago, “most of the *in-* words are settled once and for all, and have to be learned by children as wholes” (1917:140). In reality, the development of a new sense can result in the loss of the status of the
prefix as a morphological unit.

3.5.4 *Non- prefix semantics*

Marchand (1969), Funk (1971) and Lieber (2007) variously describe *non-* as a prefix which attaches to nouns (particularly proper nouns) and adjectives (almost any adjectives and proper noun adjectives). It is suggested that it is used in learned vocabulary (originally it was only in respect of legal terminology), scientific language, philosophy, religion, political history and journalistic language. Similarly, Funk states that it is “rarely used in everyday colloquial speech” (1971:372). Marchand and Mazzon describe *non-* as having become “very productive recently” (Marchand, 1969:180), and as “gaining ground in recent times” (Mazzon, 2004:111), whereas Kwon argues that *non-* displays potential, but that it is not as productive as is claimed (1997:231).

Semantically, Lieber proposes that *non-* is the most semantically restricted of all the negative prefixes for a variety of reasons (2007:114). It has neither a reversative or privative meaning and provides a purely negative sense. It is also unique in comparison to the other negative prefixes in that it is the only prefix to provide a strictly contradictory meaning (Funk, 1971:372; Lieber, 2007:114). In other words, it creates the binary antonym, and there are no shades or degrees of the quality or entity (Lieber, 2007:114). Lieber categorically states that “*(n)on-* does not attach to verbs” (2007:114), though the BNC gives one example of a *non-* prefixed verb, *non-suit* (a legal term). Certainly, however, *non-* verbs appear to be extremely rare.

As with the examples *amoral* and *immoral* discussed earlier, *non-moral* (though it only has a token frequency of 2 in the BNC) is still a further example of
the competition of the prefixes for the same environment. Indeed, an Internet search
gave over eight million hits for *amoral* and over six million for *immoral*, whereas
*non(-)moral* showed approximately 37,800,000 hits. As Kwon proposes, “there
seems to be general agreement among those who have commented on the distinction
between *non-* and *un-*: that each derivative refers to a different semantic aspect of
negation (i.e. contradictory and contrary).” (1997:197). Kwon further comments,
along with Funk, that *non-* tends to be used for classification purposes, whereas
*un-* has a tendency towards characterisation or judgement (Kwon, 1997:198; Funk,
1971:372). Kwon goes on to propose that *non-* is often used as “a conscious attempt
to change the usage” from *un-* (1997:200 & 202). More recently, Mazzon proposes
that in current English, *non-* may be chosen for ironic or euphemistic effect, for
instance, in examples such as *non-entity* and *non-standard* (2004:112).

Typically, *non-* has a more neutral sense, whereas *un-* has a tendency to carry
more critical connotations. The apparent increasing level of usage of the prefix *non-*,
as proposed by the researchers above, may well be explained in terms of this neutral
status, whereby it simply means ‘not’ but without any implied criticism, or adverse
or negative connotations. For example, *non-professional* simply means ‘not
purporting to be professional’, whereas *unprofessional* implies behaviour which
ought to be professional but is contrary to an accepted code of conduct.

(v) There is little point in getting worked up about the way different people use
words (although in my *nonprofessional* life I am quite prepared to get
worked up about people who boil lobsters alive).

(vi) Libraries were asked to say whether pre-planned formal induction, job or
development training was automatically offered to their professional or
*non-professional* staff.

It should, however, be acknowledged that a Google search of this nature, would not yield entirely
reliable figures, which are at best only approximations.
The article had suggested that, because of the actor's separation from his wife of 15 years, actress Sheila Hancock, he had behaved in an unprofessional and undignified manner, the court heard.

Examples (v) and (vi), taken from the BNC, non(-)professional (i.e. both with and without a hyphen) clearly demonstrate the neutral status of the prefix non-, whereby non- simply highlights a straightforward, non-critical contrast with professional. In contextual example (vii), however, unprofessional unquestionably has negative connotations and a criticism is definitely implied, and refers to behaviour which contrary to some accepted and expected norm. In other words, the meanings of non- and un- dictate the choice of prefix.

Apart from non- verbs being extremely rare, non- adverbs are almost as uncommon. Their scarcity is presumably to do with the neutrality of the prefix non-, which creates a straightforward binary opposition of either possessing a characteristic or not. In other words, in the case of non-, there is little need for shade, degree or modification of the attribute by the use of an adverb. Of the 3,784 non- types in the database, only 70 are adverbs. Taking a selection of the higher frequency adverbial non- types: non-inferentially (12), non-verbally (13), non(-)committally (38) and non(-)stop (115), these were examined contextually in the BNC:

(viii) He spoke noncommittally, and Lydia understood that there were things of which Beuno would not yet speak to her.

(ix) He shrugged non-committally.

(x) One whale was heard singing nonstop for at least 22 hours.

(xi) Traffic streaming non-stop over the bridge.

(xii) The regress argument is an argument that as well as the inferentially justified beliefs, there must be some beliefs which are justified non-inferentially.
(xiii) Stimulation from parents enhances the speed of learning, and great patience is required to perceive and decode the communicating done by babies and children, both verbally and non-verbally.

(xiv) Obviously he can create a better model if he knows what the interviewee looks like, how he dresses, talks and responds non-verbally.

Contextual examples (viii) to (xi), though used adverbially, are examples whereby the positive forms of the word (i.e. the prefix-free versions) tend not to occur so frequently. For example, on an Internet search for non(-)committally, approximately 358,000 results are found, whereas committally only brings up approximately 27,500 occurrences. Therefore, in examples such as these, in some respects, the non- words do not function strictly as negated words, but rather as verb modifiers, which describe how the action of the verb is performed (manner). Similar to the earlier example, infinitely, this appears to be more about the semantics of the word as a whole, rather than a comment on the semantics of the prefix.

In contextual examples (xii) and (xiii), the non- prefixed adverb co-occurs intra-sententially with the positive form of the adverb (inferentially / non-inferentially; verbally / non-verbally). As Jones suggests “(w)hen looking at ‘opposites’ within a corpus, one is immediately struck by the volume of contexts in which an antonymous pair co-occur.” (2002:25). The positive and negative versions of, in this case, the adverb are used to “signal inclusiveness or exhaustiveness of scale”; what Jones refers to as “coordinated antonymy” (2002:61). In other words, in contextual examples (xii) and (xiii), the negated non- adverb is used to show exhaustiveness in terms of encompassing all ‘justified beliefs’ and all ‘the communicating done by babies and children’ respectively. In contextual example (xiv), non-verbally is again used exhaustively in the sense that verbally is understood in the context from the verb ‘talks’, though it is not actually present. In all of these examples, the non- prefixed adverb is used either as though it were not a negated
form, or alternatively used almost in a ‘forced’ way to set up a contrast with the non-negated form to signal exhaustiveness. This provides some semantic explanation for the few adverbial non- types, which is discussed further in subsequent chapters in terms of word class distribution.

3.5.5 Un- prefix semantics

Mazzon describes un- as “the most common and productive of the negative prefixes” which, since it conveys “both complementary and contrary semantic relations” and can be attached to almost any adjective to create its negated version (2004:111). Mazzon’s view of its productivity is not necessarily supported by the current study, though this is examined further in terms of the various measures to be applied.

Many researchers (Marchand, 1969; Funk, 1971; Kwon, 1997; Lieber, 2007, etc.) suggest that the negative prefix un-, as the only native negative prefix, freely attaches to almost any adjective and is a common prefix with derived adjectives, including compound adjectives (e.g. unselfconscious). For example, the database provides many further-derived un- adjectives such as unaerodynamic, unaffectonate, unenlightening, unproblematical, etc., along with many compound adjectives such as uncrosslinked and unselfconscious. Funk also proposes that there are a number of novel compound adjectival constructions with un-, and these are demonstrated in the database with examples like unputdownable, unpickupable, unturndownable, unpigeonhole(-)able, which he believes provide “vivid evidence of the productivity of the pattern.” (1971:371)\(^{10}\). Marchand proposes that derived un- adjectives with -able are common, and this is certainly demonstrated in the database with countless examples of this combination (1969:202).

---

\(^{10}\) Horn discusses “the current proliferation of un-nouns”, in which he compiles and examines “a large corpus of novel un-nouns” in terms of prototype semantics and privative opposites. (2002: Abstract).
Furthermore, Marchand states that derived *un*- adjectives suffixed with *-like* and *-worthy* are common (1969:201-2). In the case of adjectives with *-like*, twenty-six are found in the database, with examples such as *unchildlike*, *undog-like*, *un-rabbit-like* and *unsportsmanlike* (which is also an example of a compound adjective with *un-*) (see below for a full list of adjectives of the ‘*un-*...-like’ construction in the database11). However, though Marchand proposes that derived *un*- adjectives which are suffixed with *-worthy* are common (1969:202), only five examples are found in the database: *uncreditworthy* (1), *un-newsworthy* (1), *unroadworthy* (7), *unseaworthy* (5) and *untrustworthy* (60), and as is shown, all except *untrustworthy* have particularly low frequencies.

The various linguists above also concur with the view that *un-* combines to produce both contrary and contradictory opposition. Indeed, Lieber proposes that *un-* (along with *in-*, *non-* and *dis-*) “can often express a whole range of privative, negative (both contrary and contradictory), and for that matter reversative meanings.” (2007:113).

A peculiarity of the negative prefix *un-* is its use in synthetically-negated adjectives, whereby the non-prefixed form does not exist (Marchand, 1969:202). A considerable number of these synthetic adjectives are found in the database, often with particularly high token frequencies. A sample of these from the database may include *unabashed* (62), *unbending* (20), *unending* (85), *unflinching* (36), *unmeaning* (2), *unnerving* (78), *unrelenting* (97), *unseen* (486) and *untouched* (476), though this list is by no means exhaustive.

3.5.6 Semantic factors summary

The five negative prefixes have been examined in order to determine how their semantics may influence the use of the prefixes. Based on a variety of researchers’ work (Zimmer, 1964; Marchand, 1969; Funk, 1971; Horn, 1989; Kwon, 1997; Mazzon, 2004; and Lieber, 2007), Table 3.4 summarises the semantic descriptions of the negative prefixes, which is adapted from Hamawand’s (2009) study into the semantics of the negative prefixes of English.

<table>
<thead>
<tr>
<th>Negative Prefix</th>
<th>Semantic characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a-</strong></td>
<td>Applied to animates;</td>
</tr>
<tr>
<td></td>
<td>Comments on people and their attitudes;</td>
</tr>
<tr>
<td></td>
<td>Describes divergence from some norm;</td>
</tr>
<tr>
<td></td>
<td>Outside the scope of the entity</td>
</tr>
<tr>
<td><strong>dis-</strong></td>
<td>Expresses the converse of the quality signified;</td>
</tr>
<tr>
<td></td>
<td>Comments on people and their acts;</td>
</tr>
<tr>
<td></td>
<td>Describes an evaluative quality</td>
</tr>
<tr>
<td><strong>in-</strong></td>
<td>Describes evaluation in terms of disapproval;</td>
</tr>
<tr>
<td></td>
<td>Comments on people &amp; their acts;</td>
</tr>
<tr>
<td></td>
<td>Tends to describe values &amp; abstract concepts;</td>
</tr>
<tr>
<td></td>
<td>Expresses violation &amp; infringement of the signified;</td>
</tr>
<tr>
<td></td>
<td>Describes a distinction to an extreme degree;</td>
</tr>
<tr>
<td></td>
<td>Describes an action that is impossible</td>
</tr>
<tr>
<td><strong>non-</strong></td>
<td>Applied to areas of knowledge;</td>
</tr>
<tr>
<td></td>
<td>Expresses neutral negation;</td>
</tr>
<tr>
<td></td>
<td>Implies the ‘other’ nature of an entity;</td>
</tr>
<tr>
<td></td>
<td>Tends to describe people &amp; concrete concepts;</td>
</tr>
<tr>
<td></td>
<td>Purely descriptive</td>
</tr>
<tr>
<td><strong>un-</strong></td>
<td>Implies the ‘antithesis’ of what is specified;</td>
</tr>
<tr>
<td></td>
<td>Comments on people &amp; things;</td>
</tr>
<tr>
<td></td>
<td>Tends to describe behaviour;</td>
</tr>
<tr>
<td></td>
<td>Refers to concrete entities;</td>
</tr>
<tr>
<td></td>
<td>Implies possession of deviant qualities;</td>
</tr>
<tr>
<td></td>
<td>Moves towards contradictory opposition;</td>
</tr>
<tr>
<td></td>
<td>Describes an action that is possible but difficult</td>
</tr>
</tbody>
</table>

Hamawand utilises prototype theory to morphology in order to answer questions about the multiple senses of the negative prefixes and how they are related (2009:55). “At the synchronic level, a negative prefix is made up of distinct senses
which are structured in the shape of a network. The senses are related to one another in a systematic way” (Hamawand, 2009:56). These related senses range over a continuum from centrally prototypical through less prototypical to peripheral.

3.6 Data collection and prefix research summary

This chapter has described the compilation of the database to be used in the subsequent analyses, and has also researched the historical and semantic factors which may affect the use and potential productivity of the negative prefixes. The methodology employed in the gathering of the database has been corpus-based, and the rationale for this choice has been discussed in terms of the advantages of such an approach over an intuition-based account. Though intuition-based accounts have not been discounted out of hand, it has been shown that a corpus-based perspective has advantages over this type of approach, as a corpus draws from genuine language use. Data from a corpus is both quantifiable and can also be utilised in a qualitative manner and, therefore, allows judgements to be made about intuitive theories which relate to language. Although the main focus of the analysis is on the corpus approach, this is not to the total exclusion of other methods and combinations of methods, which are also employed to support and compare with the corpus findings. The chapter has examined the reasons for the choice of the BNC as the corpus to be used in this research. The texts of the BNC were compiled from both written and spoken sources, and include data from a large variety of mediums and domains and reflect a wide range of language use, aiming to be representative of the English language as it is used in the United Kingdom.

Decisions and justifications have also been made as to which negative prefixes were to be included in the case study of this research. As opposed to
‘negative’ prefixes which incorporate additional senses to that of straightforward negation, such as ‘counteracting’ (e.g. *antifreeze*), ‘abnormality’ (e.g. *dysfunction*) or ‘wrongly’ (e.g. *misjudge*), the prefixes *a*-, *dis*-, *in*-(and its variants *il*-, *im*- and *ir*-) , *non*- and *un*- were selected as the most pertinent prefixes which best reflect pure negation. That is, the chosen prefixes strictly reflect negation.

The chapter has also described the methodology used to create the database, which has utilised Kilgarriff’s Sketch Engine software. This chapter has then recounted the rationale for, and the process of editing the database, via the strictly-applied inclusion and exclusion criteria which were drawn up for this purpose to maximise the consistency of the process. The final database employed in this research comprises a total of 9,121 distinct negatively-prefixed words (types), which can be analysed according to their frequencies and word classes, and which can be sub-divided into 343,956 actual occurrences of negatively-prefixed words (tokens).

The second part of the chapter also researched the historical and semantic factors which may affect the negative prefixes and, in the subsequent analyses, this research is drawn upon to support the findings. The following chapter not only utilises the database to analyse the negatively-prefixed words in terms of the productivity of the prefixes, but also examines the distribution of the words between the various open word classes, along with analysis of the domains which the individual prefixes may prefer. The chapter applies, adapts and develops other researchers’ methodologies, and also proposes a more refined measure of morphological productivity, with a view to determining the current productivity of the negative prefixes of English.
Chapter 4

Analysis and Discussion

In the previous chapter, the rationale for choosing a corpus-based approach was discussed, and the chapter also described how and why a database of negatively-prefixed words was compiled for use in the analyses. The previous chapter also demonstrated how other methods and combinations of methods can be used to enhance the findings of a corpus-based methodology.

From this starting point, the database of negatively-prefixed words was created from the BNC, using Sketch Engine shortcut software, to identify all negatively-prefixed words from the open word classes: adjective, adverb, noun and verb. These particular word classes were examined from the point of view that, as open classes, they are inclined towards the production of new formations; that is, they readily act as bases for productive processes. The database subsequently created was then manually edited with the use of strictly-applied inclusion and exclusion criteria which were developed and applied to ensure consistency throughout.

The aim of this thesis is to use the negative prefixes as a case study to revisit the methodologies proposed by previous researchers and to comment on the productivity of these prefixes. The purpose of the current chapter is to collect and set out the statistical information, by means of a database of all negatively-prefixed words compiled from the BNC, in order to replicate earlier studies, both pre-corpus and post-corpus, which have been put forward by a variety of researchers to measure productivity. With this in mind, the chapter applies a number of previously-tried methodologies to the case study, in order to analyse the benefits and drawbacks of
each of these methods. This chapter, therefore, gathers the evidence, re-appraises the means of measuring productivity as a whole, and moves towards the development of new more sophisticated measures, in order to determine the productivity of the negative prefixes.

To begin, the database of negatively-prefixed words are analysed to identify patterns of usage among the various prefixes. The literature referred to in Chapter 2 is evaluated with respect to the corpus evidence, whereby the database provides information for use in a variety of relevant analyses, which involve the various formulae proposed in these studies. The chapter examines a number of variables which have previously been used to measure productivity to determine which variable or combination of variables constitutes the most reliable indicator. Underpinning many of these measures is the idea that the presence of a neologism which incorporates a given affix will reflect the productivity of that affix. However, what has been in question to date is the facility to be able to identify neologisms. This chapter, therefore, pinpoints negatively-prefixed neologisms in order to ascertain which of the various negative prefixes is the most productive. Harald Baayen, in conjunction with other researchers in the late 1980s through to the mid 1990s, proposed that hapax legomena (referred to as hapaxes), words which occur once in a given corpus, could potentially be equated with neologisms, and that to examine the hapaxes in a corpus was, for the most part, equivalent to examining the newly-created words of a language. Baayen and his co-authors have, therefore, promoted the presence of hapax legomena as a means of identifying new formations in the language. This chapter identifies the hapax legomena in the database for each of the negative prefixes, examines them in greater detail than has previously been undertaken, and considers their relevance in measuring productivity. However, in
addition, the chapter works towards the development of a more refined method for the identification of neologisms, and gauges the importance of hapaxes in relation to their possible status as neologisms.

This chapter also investigates the negative prefixes in terms of their word class distribution, along with an analysis of the database to ascertain which domain each negative prefix favours. As a whole, the chapter deals with the procedural aspects of the research: the collection of the statistical evidence by means of reporting frequency information. It also discusses and interprets the development of new, more-refined methods of measuring productivity based on the implications of this evidence.

Broadly speaking, the study replicates the 1991 measures proposed by Baayen and Lieber but uses a larger and more recent corpus, and though the research has chiefly focused on corpus-based work of this type, non-corpus methods have also been drawn on, such as those applied by Plag (1999) and Bolozky (1999). As outlined in Chapter 2, Baayen and Lieber examine a number of affixes, though these are mostly suffixes, and they apply productivity measures to each. For instance, they use a hapax/token measure to calculate the productivity of the noun-forming suffix \textit{-ness}, which has 77 hapaxes in their corpus and a token frequency of 17,481, and gives a productivity quotient of 0.0044. Baayen and Lieber’s intuition, that \textit{-ness} is a productive suffix, is borne out by this relatively high figure, when compared to other suffixes which they examine. By comparison, for example, Baayen and Lieber examine the adjectival suffix \textit{-ous}, which shows 13 hapaxes in their corpus and 21,861 tokens. Baayen and Lieber argue that, although \textit{-ous} has a higher token frequency, it has fewer hapaxes and, therefore, as a result, is less productive than the
suffix -ness with a measure of 0.0006 (1991:819 and 821). Typically, however, Baayen and Lieber (1991) focus on class-changing affixes, whereas the current study, in examining negative prefixes, centres on affixes which are class-maintaining. Nevertheless, Baayen and Lieber’s measures, which are replicated and critiqued by the current research, are an established methodology, which demonstrate the possibility of applying numeric values to particular word formation processes.

Much of the previous corpus-based research into morphological productivity, such as that by Baayen and his colleagues, has centred on the relevance of the notion of hapax legomena, which have been viewed as a crucial factor in a number of their measures, and one of the key research questions of the current thesis has been to examine the value of hapaxes in productivity measures. The chapter will show, for example, that while hapax legomena can be treated as a vehicle to access neologisms, they are not neologisms per se, which brings Baayen and Lieber’s (1991) productivity measures into question, since, rather than taking them at face value, hapaxes have been shown to require more detailed investigation if they are to be used in a measure of productivity. The current chapter also discusses the importance of neologisms to a study of productivity, and the correlation between hapax legomena and neologisms are used to develop a more fine-grained measurement of morphological productivity. In addition to this, the current chapter further examines the effect that inclusion in a particular word class may have on the productivity levels of the negative prefixes.

Productivity remains a complex issue, but the current chapter aims to resolve some of the issues which have been raised. The evidence presented in the previous

---

1 Baayen and Lieber (1991) do in fact examine the negative prefixes in- and un- via the same measure, though the findings are particularly counter-intuitive. The calculations for un- show a lower level of productivity than for the suffix -ous with a figure of 0.0005, and for in- the productivity is less again at 0.0004, but is not, as would be expected, particularly lower than for un-.
chapter is, therefore, interpreted more fully and the broader questions set out at the start of this thesis are answered by providing the relevant evidence. Prior to the use of electronic corpora, measures, such as those proposed by Aronoff (1976), were by necessity based on dictionary counts, and were, at best, subjective type frequencies. With the event of corpora, however, actual frequencies (in other words, objective type and token frequencies) were able to be measured, which demonstrate the specific occurrences of a word type and also provide an indication of its frequency of use in the language, both of which may be viewed as valuable in a measure of productivity. Therefore, in addition to investigating hapaxes more closely, the current chapter further examines type and token frequency and, with a view to developing existing measures, works towards new, more refined measures of morphological productivity.

The measures concerned with type frequency, token frequency and the type/token ratio are compared first, and the benefits and limitations of each measure are discussed. Hapaxes are an important aspect in determining the productivity of a word formation process, as a means rather than as an end, and the hapax/type and hapax/token measures are further discussed on this basis. However, this chapter then goes on to propose new formulae, which make use of the neologism:hapax ratios, while also factoring in the effect of word class on productivity.

4.1 Replicating methods for measuring productivity

One of the most central issues [....] to arrive at a less naive and a better articulated conception of morphological productivity is to come to grips with the factors determining, influencing and/or affecting the property of morphological processes to produce new words.

(Van Marle, 1992:151)
Language productivity has always been a controversial issue and linguists do not always agree on how the output of productivity should best be measured. As described in the previous chapter, ‘types’ refer to the number of different words which occur in a corpus; in this case, the individual words which are formed that incorporate a particular prefix. Thus, disagreement, non-existent and uninhibited are negatively-prefixed types which occur in the BNC and, therefore, in the database.

‘Tokens’, on the other hand, are the actual number of occurrences of a particular type; the number of times that a particular type appears in the BNC/database. Therefore, the negatively-prefixed type disagreement occurs 1,158 times in the BNC/database, non(-)-existent appears 431 times in the BNC/database and uninhibited occurs 101 times. ‘Hapax legomena’, as individual words which occur only once in a corpus, are types which have a token frequency of only one. For example, areligious, discommunication, inappropriacy, non-academically and unegalitarian are all hapax legomena which appear only once in the BNC/database.

Using the negative prefixes as a case study, this chapter revisits five different, previously-tried methods for the measurement of productivity, which utilises these three variables (types, tokens and hapaxes), either singly or in combinations. Specifically, these five methods are:

- ‘Type’ method
- ‘Token’ method
- ‘Type/token’ method
- ‘Hapax/token’ method
- ‘Hapax/type’ method
4.1.1 The ‘type’ method

In the context of productivity studies, type frequency refers to the number of distinct words which, at some point, were coined using a particular word formation process. In using this method, one essentially identifies how many different words in the language currently make use of a particular affix. The type frequency approach to productivity prioritises breadth of coverage, whereby the important issue is the number of different words, not how common these individual words are in the language. As the following section reports, the main advantage of a type-counting measure is that it can be based on the evidence of dictionaries or even intuition; for example, (actual) type frequency can be compared to potential type frequency in order to calculate a productivity ratio, as the next section explains.

4.1.1.1 Aronoff’s (1976) type/potential type method

Though Tottie (1980), Baayen and Lieber (1991), Baayen (1992), Frauenfelder and Schreuder (1992), Van Marle (1992), Schreuder and Baayen (1994), Baayen and Renouf (1996), Kwon (1997), Plag (1999) and Plag et al (1999) have all, to some extent, discussed the relevance of type frequency in terms of measuring productivity, it is typically the studies that predate widespread corpus availability which have concentrated most on type measures. For example, Aronoff (1976) proposed that the productivity of a word formation rule (WFR) can be determined by the use of a type-based measure “by counting the number of possible bases for the rule” (possible types), and then counting the number of “actually occurring words formed by that rule” (actual types) (36). A ratio is then taken of the two figures, which can then be compared with the ratios for other word formation rules in terms of a measure of their productivity. Thus:
Productivity = \frac{\text{Number of actual types}}{\text{Number of possible types}}

So, for example the productivity of *un-* among adjectives would be expressed as:

\[
\text{Productivity} = \frac{\text{un} + \text{Adj}}{\text{Adj}}
\]

Determining the number of potential types a particular affix may give rise to is problematic, since being able to estimate the number of possible types which may arise is a potentially infinite, and thus impracticable, task. For example, the suffix *-able* relatively freely attaches to an array of nouns and verbs to create adjectives (*comfortable, breakable*) and is, as a result, considered a productive affix. In turn, the prefix *un-*, also considered a productive affix, can then be attached to these adjectives suffixed with *-able* to create negative adjectives (*uncomfortable, unbreakable*). Determining whether it is the suffix *-able* or the prefix *un-* which is the more productive element is a near impossible task, as indeed is identifying the total number of potential combinations that utilise these two affixes. A measure which focuses solely on type frequency and potential output is, therefore, seemingly impossible to calculate.

These methodological weaknesses did not pass unnoticed in the literature. For example, Baayen and Lieber note that one of the problems of this possible/actual type ratio is that “Words are not ‘actual’ or ‘existing’ in any objective sense.” (1991:802), and quantifying any list of words, whether from someone’s mental lexicon or from a dictionary, is inevitably subjective. Likewise, as Aronoff himself notes, each time a new word formation occurs it is not registered in a list, yet
“(u)nless all new words are listed, we have no effective procedure for computing the ratio of existing to possible words” (1976:36).

4.1.1.2 Bolozky’s (1993; 1999) dictionary method

Bolozky (1993; 1999) also advocates type-based methods for measuring productivity, most notably his dictionary-based measure (1999:4). Essentially, dictionaries are lists of types and, whilst dictionary-based methods cannot be totally relied upon, since they depend on the lexicographer’s judgement as to inclusion and exclusion and are, therefore, to some extent, subjective, Bolozky believes that they have their purpose. He proposes that comparisons can be made between an earlier and a later version of the same dictionary, which “allows one to determine which words constitute recent innovations, and may thus point to recent tendencies in word formation.” (1999:5). Bolozky proposes a ratio which expresses the total number of new words which incorporate a particular affix which have been added by the time of the later dictionary, set against the total number of words which include that affix in the earlier version of the dictionary. (1999:5). The formula is expressed as follows:

\[
\frac{\text{Total number of additional words incorporating affix}}{\text{Total number of words incorporating affix}}
\]

Hulse (2002) examines the negative prefixes in terms of Bolozky’s measure with the aid of earlier and later versions of the *Concise Oxford Dictionary: 1976 (COD76)* and 1996 (COD96). Although Bolozky focuses on suffixes, Table 4.1 below demonstrates Bolozky’s measure applied to the negative prefixes.
Table 4.1. Types (in 1976), Additions (by 1996), and Proportion of New Types (after Hulse, 2002, based on Bolozky, 1999)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>non-</td>
<td>34</td>
<td>79</td>
<td>2.324</td>
<td>232.4%</td>
</tr>
<tr>
<td>2</td>
<td>a-</td>
<td>53</td>
<td>33</td>
<td>0.623</td>
<td>62.3%</td>
</tr>
<tr>
<td>3</td>
<td>dis-</td>
<td>31</td>
<td>16</td>
<td>0.516</td>
<td>51.6%</td>
</tr>
<tr>
<td>4</td>
<td>un-</td>
<td>1,079</td>
<td>183</td>
<td>0.170</td>
<td>17.0%</td>
</tr>
<tr>
<td>5</td>
<td>in-</td>
<td>330</td>
<td>6</td>
<td>0.018</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

What is interesting to note about the figures in Table 4.1 is that, while un- prefixed words have, by a huge margin, the largest number of inclusions in the COD 1976 and the number of additions by the COD 1996 is again the largest, the actual percentage increase from the COD 1976 to the COD 1996 is one of the lowest. On the other hand, non- prefixed words in the earlier version of the COD have one of the lowest numbers of inclusions, yet the increase by the 1996 version is well over 200%.

As discussed earlier, a number of linguists, notably Bolozky (1993; 1999) and Plag (1999), promote the usefulness of dictionaries in measuring productivity. What most dictionaries do not provide, however, is any indication of how often a word is used in a language, though some, such as the OED, may specify if a word is currently rare or even likely to become obsolete. Moreover, until relatively recently, most dictionaries relied heavily on lexicographers’ instincts as to the inclusion of a new word. Despite contemporary dictionary-compilers being more likely to check the corpus frequency of potential new additions, still no specific criteria exist to determine the inclusion or exclusion of a new word. As Lieber (2010) points out, “makers of dictionaries – lexicographers – are of course human; what gets into dictionaries has historically been subject to the individual foibles of lexicographers,
not to mention the mores of society” (13), and she goes on to state that to decide which words should be included in a dictionary:

....is a judgment call, often based more on practical considerations – the size of the dictionary, its intended audience – than on strict linguistic principles. All lexicographers face this conundrum, and each one makes a slightly different decision.

(2010:26-7)

It should be noted from Table 4.1 that whilst un- has the largest number of entries by the 1996 version of the dictionary (COD) with 1,262 entries, this does not necessarily indicate that un- more readily attaches to base words. It should also be noted that suffixes are more easily dealt with by a dictionary, since they can be listed within the base word entry. Prefixes, on the other hand, have to be alphabetically separated from their base words and have to be listed as ‘headwords’. Prefixes are, therefore, less easy to deal with and lexicographers must decide on their inclusion or exclusion on a case-by-case basis. As the Collins Concise Dictionary (1995) states “Prefixes (e.g. in-, pre-, sub-) [....] have been entered as headwords if they are still used freely to produce new words in English”. Therefore, decisions are made on the productivity of prefixes before they gain inclusion in a dictionary. One must conclude, therefore, that dictionary-based type methods such as Bolozky’s, whilst useful in some respect, are too reliant on subjective inclusion/exclusion decisions to be considered robust.

4.1.1.3 Type Counts for Negative Prefixes

Although type-based methods can be used to shed some light on productivity, a straightforward type-counting measure alone remains a relatively crude indicator.
Aronoff’s (1976) position is more a statement of principle than an actual procedure (albeit one that points towards future, empirical methods with some foresight). Bolozky (1999) attempts to pin down this principle by using dictionary-based techniques, but encounters the problems inherent in any such method, namely that the results rest too heavily on the whims of individual lexicographers.

Nevertheless, type information is valuable to studies of morphological productivity because, as subsequent sections show, the number of different words that make use of a given affix in a language is seen as a key variable in more sophisticated measurements of productivity. Table 4.2, therefore, shows the distribution of negatively-prefixed types in the BNC, with the prefixes un- and non- respectively which demonstrate considerably higher number of types than, for instance, the a- prefix.

<table>
<thead>
<tr>
<th></th>
<th>Number of types per prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>un-</td>
<td>3,836</td>
</tr>
<tr>
<td>non-</td>
<td>3,784</td>
</tr>
<tr>
<td>in-</td>
<td>922</td>
</tr>
<tr>
<td>dis-</td>
<td>429</td>
</tr>
<tr>
<td>a-</td>
<td>150</td>
</tr>
<tr>
<td>Total number of types</td>
<td>9,121</td>
</tr>
</tbody>
</table>

In itself, Table 4.2 is in no way indicative of the productivity of the negative prefixes. The concept of ‘type’ does not account for the number of times these individual words are encountered or in what circumstances; it merely lists their existence. Furthermore, Table 4.2 says nothing about when the 9,121 word types were coined or whether the process by which they were formed remains an active
one. However, the relevance of this information becomes clearer as the chapter progresses.

4.1.1.4 Discussion of the ‘type’ method

As described above, the ‘type’ method involves a straightforward count of the number of different types created by a particular process, and it has been claimed that by adding the total number of types created by a particular process is an indication of the productivity of that process. Typically, the idea of type-frequency counting belongs to the pre-corpus era, since token frequency information was not available at the time, though almost all proponents of type-frequency measures subsequently reject them as too simplistic, as they deal with either potential types or dictionary type counts. As Van Marle observes “many morphologists tend to take the stand that the property of morphological processes to underlie new words is determined both by forces pertaining to the competence and by forces relating to language use” (1992:151). He claims, therefore, that, as a result, simple type-counting methods came to be rejected in favour of methods which involved more than one variable. Aronoff’s (1976) measure, for example, involved predicting possible types and counting actual types, though clearly, quantifying possible types is not practicable. A dictionary-based (and therefore type) measure, as proposed by Bolozky (1993; 1999), was also examined in the previous chapter, though it was proposed that the inclusion of words in dictionaries is not only subjective, as it is reliant on the lexicographer, but furthermore, dictionaries do not necessarily list every example of a word which incorporates a particular affix. For example, dictionaries do not list every possible word prefixed with un-,
possible further-derived word that involves un-. So, for instance, unbouncebackability is not likely to appear in a dictionary.

Bolozky’s (1993; 1999) dictionary measure compares the additional types in a later version of a dictionary compared to an earlier version of the same dictionary. According to this measure, Bolozky proposes that “(t)his method allows one to determine which words constitute recent innovations, and may thus point to recent tendencies in word formation” and subsequently to be able to identify patterns in the use of affixes (5). This measure was also applied to the negative prefixes by Hulse (2002) and non- was found to be by far the most productive of the negative prefixes, whereas in- was shown to be the least productive.

Despite the shortcomings of type-counting/dictionary methods, it is interesting to note that the results of Bolozky’s measure when applied to the negative prefixes has a striking similarity to the productivity measures to be examined in greater detail in the following sections of this chapter. Nevertheless, a straightforward list of types can, at best, merely indicate the existence of a word in an abstract way, which marks its potential use; but alone, raw data in respect of types serve no practical purpose with regard to measuring productivity, and types alone fail to signal how and how often a word is used in language. As Plag (2003) notes:

There is one quantitative measure that is probably the most widely used and the most widely rejected at the same time. According to this measure, the productivity of an affix can be discerned by counting the number of attested different words with that affix at a given point in time. This has also been called the type frequency of an affix. The severe problem with this measure is that there can be many words with a given affix, but nevertheless speakers will not use the [affix] to make up new words.

(2003:52)

2 Used by the character Peter Barlow in the television series, Coronation Street (14/09/09) to describe his then state of mind.
For these reasons, Table 4.2 merely demonstrates that \textit{un-} and \textit{non-} have the largest number of types; it is not in itself an indication of the productivity of these two prefixes. Thus type, and therefore dictionary, measures can, as seen above, only serve to verify other corpus-based measures which have been applied, but do not in themselves engage with the notion of productivity in the sense discussed by this thesis.

\subsection*{4.1.2 The ‘token’ method}

Whereas type frequency lists the number of individual words created by a specific process, tokens are the number of actual occurrences of a particular type. For example, the word \textit{atypical} is an example of a negatively-prefixed word type which appears in the \textit{BNC}, and occurs 144 times. \textit{Atypical}, therefore, has a token frequency of 144 in the database. ‘Token’ methods refer to a simple count of the number of times a particular type occurs in a corpus. Therefore, the notion of type frequency provides a breadth of coverage approach, and records the number of different words coined that utilise a particular process, whereas token frequency, on the other hand, provides a depth of coverage approach, and records how commonly the types are actually utilised in the language. Tokens, therefore, could be said to provide a reflection of actual usage.

While a corpus can be searched to produce a list of types, it can also provide information on token frequency. Whereas type-counting methods were often based on potential, intuition and subjective evidence from dictionaries, token frequencies are based on actual occurrences, from genuine language data. Although still in the pre-corpus era, and initially having concerned himself with type frequencies, by
1983, Aronoff had also begun to look towards the relevance of token frequencies in measuring productivity. Various researchers (Tottie, 1980; Anshen and Aronoff, 1988; Baayen and Lieber, 1991; Baayen, 1992; Baayen and Renouf, 1996), around this time and since, have concerned themselves with the output of productive affixes; in other words, with token frequency, though it has been with the advent of electronic corpora that methods which involve tokens as a variable have come to the fore.

Anshen and Aronoff (1988) in particular, examine the notion of the relationship between token frequency and productivity (642), though their particular focus is on the relationship between the base word and the derived word. They argue that while some words which comprise a base word and one or more affixes, in other words, morphologically-complex words, are stored in the mental lexicon, others are constructed as required, and that to retrieve a complex word, we simultaneously try to locate the whole word, whilst also trying to build the complex word using its components and the rules of the language (642). Anshen and Aronoff (1988) attempted to find frequency-based evidence for Aronoff’s (1976) study (Baayen and Lieber, 1991:806). Their research involved a production test, whereby participants were asked to produce a list of words which contained the suffixes: -ibleness, -ibility, -iveness, -ivity, -ionary and -ional. The test showed that the participants were more likely to produce nonce formations in -ibleness and -iveness than with the other formations (Baayen and Lieber, 1991:805). In their experiment, Anshen and Aronoff (1988) propose, therefore, that -ness formations are consequently constructed as required in the mental lexicon, whereas forms which incorporate -ity are stored in the mental lexicon as whole constructions. From this, Anshen and Aronoff conclude that
the noun-forming suffix -ness is more productive than the suffix -ity, and go on to make a connection, in fact an assumption, that all constructed forms must, therefore, be productive, whereas formations which are stored whole are not (1988:642). However, it is questionable whether there is any logical connection between productive affixes such as -ness (those with a low type/token ratio), and word formations which are constructed in the mental lexicon. It may simply be the case that -ness is more productive due to the large number of base words with which it can combine. “Moreover, it is rather counterintuitive to claim that no formations in -ness are stored [...] since it is unlikely that concepts like forgiveness or effectiveness are reinvented for each successive instance of use” (Baayen and Lieber, 1991:806).

The Anshen and Aronoff production-test experiments are based on Aronoff’s (1982), seemingly incongruous, data taken from Walker’s Rhyming Dictionary (1936) and a one-million word corpus created by Kučera and Francis in 1967. In Aronoff’s (1982) calculations, “he arrives at a mean frequency of 9.565 token/type for the 23 types [in] Xivity”, whereas there is “a mean frequency of 0.641 for the 103 types in Xiveness” (Baayen and Lieber, 1991:804).

As Baayen and Lieber propose, the notion that words which have been derived through productive processes are never stored in the mental lexicon, but are always constructed, is “too simplistic” (1991:806). Anshen and Aronoff’s findings suggest that higher frequency word formations in Xivity “block the corresponding formations in Xiveness [...] while at the same time the rule-generated formations in -iveness block ‘access to and thus the existence of a lexically based -ivity form’ (1988:653, cited in Baayen and Lieber, 1991:806). Baayen and Lieber (1991) state that this “cannot be correct” either “logically since, as pointed out by van Marle

---

1 They later question this notion in Anshen and Aronoff (1989).
(1985), words cannot at the same time block and be blocked” and empirically, as there is no reason to suggest that -ness formations are never stored. (806). Baayen and Lieber also go on to argue that “the relation between the frequencies of the base and derivatives are not particularly relevant to the study of productivity” (1991:807).

### Table 4.3. Negative Prefixes – Total number of tokens for database types

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Number of tokens per prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>un-</td>
<td>170,984</td>
</tr>
<tr>
<td>in-</td>
<td>108,791</td>
</tr>
<tr>
<td>dis-</td>
<td>38,368</td>
</tr>
<tr>
<td>non-</td>
<td>22,466</td>
</tr>
<tr>
<td>a-</td>
<td>3,347</td>
</tr>
<tr>
<td>Total number of tokens</td>
<td>343,956</td>
</tr>
</tbody>
</table>

Table 4.3 shows the distribution of tokens across the negative prefixes of the database. Initial observation shows that the prefixes un- and in- have, by a considerable margin, the highest token frequencies of all the negative prefixes, and the prefix a- has the least number of tokens. Again, however, as with type frequency, token frequency does not in itself constitute an indication of productivity, since high numbers of tokens does not indicate current productivity. Likewise, token frequency does not provide any information as to when the words involved were first coined, or whether the process involved still remains fertile.

Baayen and Lieber (1991) make use of token frequency in their studies of productivity, but specifically examine only the lowest token frequencies: the hapaxes (801). Commenting on Baayen’s (1989) study, Van Marle states that “Baayen’s main claim is that productivity and frequency are not only closely correlated but that the study of frequency does contribute to our understanding of the phenomenon of morphological productivity” (1992:152). Plag notes that:
Counting derivatives can nevertheless be a fruitful way of determining the productivity of an affix, that is, if one does not count all derivatives with a certain affix in use at a given point in time, but only those derivatives that were newly coined in a given period, the so-called neologisms.

(2003:52)

It would appear that few, if any, researchers have focused solely on token frequency as a measure of productivity. Token frequencies are used more as a means to access the usage of newly-created words, rather than as a variable in their own right. Despite the fact that token frequencies are compiled from authentic language data, a simple token-counting measure is also a somewhat naive approach, since, along with type-counting measures, individually they can at best only provide a single-dimensional perspective. The presence of a high number of tokens is not necessarily an indication of a high degree of productivity, nor does low token frequency necessarily indicate a lack of productivity. As Anshen and Aronoff (1989) concede:

....the absolute number of words of a given form existing in English need not reflect the current productivity of an affix. It is possible to have a large number of words of a given form, none of which have been coined in the last three centuries.

(199)

What is of significance, however, is that Tables 4.2 and 4.3 do not appear to mirror one another. In other words, just because an affix has a large number of types, it does not follow that the total number of tokens will be comparably high. This observation, which has formed the basis for several approaches to measuring productivity, is explored in more depth in the next section.
4.1.2.1 Discussion of the ‘token’ method

As discussed above, ‘token’ methods refer to a simple count of the number of times a particular type occurs in a corpus. Tokens could be said to provide a reflection of actual usage, and thus, with the token frequency approach, depth of occurrence is prioritised.

In Anshen and Aronoff’s (1988) production-test research, their claim is that all words with productive affixes (such as the suffix -ness) are constructed as required in the mental lexicon, whilst words with unproductive affixes (such as -ity) are stored already incorporated into derived words. Making a connection between this idea and productivity, it is their belief that, as a consequence, productive affixes have the higher token frequencies, since productive affixes are actively used to coin new words, whereas unproductive affixes simply ‘exist’ in the already-derived word but are not actively used in new constructions (642). However, as Baayen (1992) points out, “Anshen and Aronoff (1988) recognize the relevance of token frequencies, but evade the problem [....] by stipulating that productive formations are not stored at all”, and Baayen goes on to state that “token frequencies do not lend themselves for a formal representation” such as that which Anshen and Aronoff propose (138). Although Table 4.3 above shows that the prefixes un- and in- have by far the highest token frequencies, again, as with type frequencies, this in no way conveys any suggestion that these two prefixes are the most productive or that low token frequency, as shown by the prefix a-, indicates a lack of productivity. Individually, and in isolation, neither raw type frequency nor raw token frequency has much to say about the concept of productivity. They do not indicate the presence of new words.
4.1.3 The ‘type/token’ method

Once it became possible to record type and token frequency via the use of corpora, researchers began to look towards a measure of productivity which involved both variables; a measure which takes a two-dimensional approach. One key measure is the type/token method, or the inverse calculation (token/type) which expresses mean token frequency per type.

As discussed in Chapter 2 (Section 2.1), Aronoff’s early work (1976; 1980) concentrated on the raw frequency of tokens. Aronoff believed that a large number of tokens per type reflected high productivity. Thus, according to his view, taking the prefixes un- and non- as examples, although un- and non- have a similar number of types, un- would be considered by far the more productive of the two prefixes because it has approximately eight times as many tokens per type as non- (See Table 4.4).

Baayen and Lieber (1991), however, approach the measure from an entirely different perspective, and propose “the inverse of Aronoff’s index of productivity” (838), arguing that it is not how many tokens per type that is important, but in fact how few. Such thinking triggered a number of studies, such as Baayen (1992), Baayen and Renouf (1996), Plag (1999) and Bolozky (1999), all of which made use of a type/token ratio. These studies followed Baayen and Lieber in assuming that lower mean token frequency is associated with higher productivity, whereas higher mean token frequency correlates with lower productivity (1991:809), which was in direct contrast with the ‘index of productivity’ proposed by Aronoff (1976; 1980).

Table 4.4 below shows how the individual type and token frequencies for the negative prefixes in the BNC can be used to create type/token ratios.
Table 4.4. Negative Prefixes – types, tokens and type/token figures

<table>
<thead>
<tr>
<th></th>
<th>No. of TYPES</th>
<th>No. of TOKENS</th>
<th>Mean Type Frequency</th>
<th>TYPE/TOKEN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>non</strong>-</td>
<td>3,784</td>
<td>22,466</td>
<td>5.94</td>
<td>0.168</td>
</tr>
<tr>
<td><strong>a</strong>-</td>
<td>150</td>
<td>3,347</td>
<td>22.31</td>
<td>0.045</td>
</tr>
<tr>
<td><strong>un</strong>-</td>
<td>3,836</td>
<td>170,984</td>
<td>44.57</td>
<td>0.022</td>
</tr>
<tr>
<td><strong>dis</strong>-</td>
<td>429</td>
<td>38,368</td>
<td>89.44</td>
<td>0.011</td>
</tr>
<tr>
<td><strong>in</strong>-</td>
<td>922</td>
<td>108,791</td>
<td>118.00</td>
<td>0.008</td>
</tr>
<tr>
<td><strong>All five negating prefixes</strong></td>
<td><strong>9,121</strong></td>
<td><strong>343,956</strong></td>
<td><strong>37.71</strong></td>
<td><strong>0.027</strong></td>
</tr>
</tbody>
</table>

Table 4.4 is ranked according to the highest type/token ratio calculated. The prefix *non*- has 3,784 types and 22,466 tokens, which gives a mean type frequency of 5.94 and a type/token ratio of 0.168. *Un-*-, on the other hand, has a very similar number of types (3,836), but many more tokens (170,984), which results in a mean type frequency of 44.57 and a type/token ratio of 0.022. According to Baayen and Lieber (1991), therefore, this suggests that *non*- has a higher level of productivity than *un*-.

4.1.3.1 Discussion of the ‘type/token’ method – Measure A

The logic behind Baayen and Lieber’s (1991) argument is that productive affixes will be associated with words that, on average, occur less in a corpus because those affixes constantly generate new items, and therefore keep the mean token frequency down. With measures that involve both of these variables, it is possible to comment in some way on productivity. This method (Measure A) is, therefore, the first of three which are discussed in greater depth.

Neither of the first two methods investigated, which separately use type and token counting, are considered useful, since neither method says anything about current language innovation. A method which involves both type and token frequencies, however, comments on both the occurrence of a new type and the number of times that type is used in the language. According to the application of Baayen and Lieber’s method (1991:809), the lowest mean type frequency and the
highest type/token ratio is *non*-, which makes *non*- the most productive of the five negative prefixes. By the same method, *in*- is the least productive of the prefixes, with the highest mean type frequency, and therefore, the lowest type/token ratio. This measure is returned to in Section 4.1.6 and further discussed in Section 4.1.6.1.

4.1.4 The ‘hapax/token’ method

In 1989, Baayen introduced a new variable which he utilised in his measures of productivity; the notion of hapax legomena, which are the types that occur only once in a corpus and, therefore, have a token frequency of one. Baayen claims that the number of hapaxes for a particular process should correlate to the number of neologisms for that process, and that the number of hapaxes can, therefore, be viewed as an indication of the productivity of the process. Initially, Baayen (1989) proposed that productivity can be measured by the calculation whereby the number of hapaxes for a given process are divided by the total number of tokens for that process, and this is taken up later by Baayen and Lieber (1991) who argue that this measure “expresses the rate at which new types are to be expected to appear” (809).

Baayen and Lieber’s interest in hapaxes is a logical extension of the view that low mean token frequency is indicative of high productivity, since hapaxes are the ultimate in low token frequency. Their hapax/token measure is, therefore, seen as a guide to the number of neologisms which could be expected to occur (813), and they provide a formula for measuring productivity based on this relationship (809).

\[
\text{Productivity} = \frac{\text{Total number of hapaxes with given affix}}{\text{Total number of tokens with given affix}}
\]

Baayen and Lieber’s belief is that highly productive affixes are characterised by a high proportion of hapaxes, in comparison to all of the tokens for that particular affix
They claim that:

...a measure of productivity based on the token frequencies of types, specifically on the number of hapax legomena for a given affix in a corpus, comes very close to according with our intuitions about productivity.

(1991:801)

Baayen and Lieber’s (1991) hapax/token measure is, therefore, applied to the negative prefixes and is demonstrated in Table 4.5 below. It calculates the hapax/token ratio of the negating prefixes in line with Baayen and Lieber’s measure (1991:809). The highest figure, by far, is for *non*, which shows a hapax/token ratio of 0.089. *Dis*- and *in*- have the lowest ratios with 0.003 and 0.002 respectively. According to Baayen and Lieber’s (1991) measure, this would suggest that *non-* is the most productive of the negative prefixes, while *dis-* and *in-* would be classed as the least productive of the negative prefixes. These findings are discussed further in the next section.

<table>
<thead>
<tr>
<th>Prefix</th>
<th>No. of HAPAXES</th>
<th>No. of TOKENS</th>
<th>HAPAX/TOKEN</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>non-</em></td>
<td>1,998</td>
<td>22,466</td>
<td>0.089</td>
</tr>
<tr>
<td><em>a-</em></td>
<td>50</td>
<td>3,347</td>
<td>0.015</td>
</tr>
<tr>
<td><em>un-</em></td>
<td>1,468</td>
<td>170,984</td>
<td>0.009</td>
</tr>
<tr>
<td><em>dis-</em></td>
<td>112</td>
<td>38,368</td>
<td>0.003</td>
</tr>
<tr>
<td><em>in-</em></td>
<td>189</td>
<td>108,791</td>
<td>0.002</td>
</tr>
<tr>
<td>All five negating prefixes</td>
<td>3,817</td>
<td>343,956</td>
<td>0.011</td>
</tr>
</tbody>
</table>

4.1.4.1 Discussion of the ‘hapax/token’ method – Measure B

According to Bolinger (1948:18), a measure of productivity should convey the “readiness with which an element enters into new combinations” (cited in Baayen

---

It should be noted that, between pages 809 and 810 of Baayen and Lieber’s (1991) article, they appear to switch between their discussion of types and tokens, and their explanation is not always completely explicit.
and Lieber, 1991:809). Subsequently, Baayen and Lieber’s hypothesis also takes into account hapaxes as the means of assessing the probability of new types occurring, along with token frequency. The measure is based on the same logic as above, namely that highly productive affixes are associated with lower frequency items. However, rather than focus on mean token frequency, the measure turns to the ultimate in low frequency phenomena, the hapax. Essentially, this gauges the proportion of times that the occurrence of an affix is part of a hapax, relative to its total use in the corpus.

The hapax/token relationship for the negative prefixes is demonstrated by the replication of Baayen and Lieber’s chart (1991:819). In this case, the vertical axis represents the number of types which are hapaxes and the horizontal axis shows the hapax/token ratio. This is demonstrated in Figure 4.1 below.

Figure 4.1. Hapax/token ratio
(after Baayen and Lieber’s Global Productivity measure, 1991:819)

Baayen and Lieber (1991) propose that the more productive affixes have higher values on both axes of the chart (in this case, hapax numbers and token frequencies).
The chart illustrates that *non-* scores highly on both axes which, according to Baayen and Lieber, classes it as a ‘productive’ affix\(^5\), with *a-*, *un-* and *in-* classed as considerably less productive affixes according to this measure, although *un-* does show a larger number of hapaxes than the other three prefixes. Whilst different negative prefixes, such as *non-* and *un-* for instance, may have similar numbers of types (3,784 and 3,836 types respectively), their token frequencies are altogether dissimilar (i.e. 22,466 and 170,984 tokens respectively). Baayen and Lieber (1991) propose that affixes with lower token frequencies are far more likely to be used to create neologisms and, therefore, that lower token frequencies indicate higher productivity (810). *Non-*-, for example, has 1,998 hapaxes in the database, whereas *un-* has 1,468 hapaxes. In other words, the vast majority of *non-* prefixed types have very low token frequencies and *non-* is, therefore, more likely to be used in the creation of a neologism. By this method, *non-* is again the more productive of these two negative prefixes.

### 4.1.5 The ‘hapax/type’ method

The type/token and the hapax/type methods examined in the last two sections have, as Baayen and Lieber suggest, “focused on productivity in the strict sense, namely, as the aspect of potentiality of word-formation rules.” (1991:817). However, in addition to the methods described to date, Baayen and Lieber (1991) propose a further measure of productivity, whereby they consider an affix to be productive if a

---

\(^5\) In Baayen and Lieber’s Global Productivity chart (1991:819), the distinction between types and tokens is again unclear. They suggest that one axis relates to types but in fact it refers to frequency, which implies tokens. Baayen and Lieber’s chart actually includes the prefixes *un-* and *in-* in their sample of affixes, but the chart appears to show that *in-* is more productive than *un-*.. For their research, Baayen and Lieber (1991) make use of an 18-million word database from the Dutch Centre for Lexical Information in Nijmegen, CELEX, version E1.0, which has been compiled on the basis of the Cobuild Project Corpus from the University of Birmingham (1991:803).
high proportion of its types are hapaxes. Baayen and Lieber refer to this as a measure of ‘global productivity’, and claim that:

...the notion of productivity can also be understood in a less-specific way when the numbers of different types are the main object of interest. Although $V$ [total number of types] is, by itself, not a measure of potentiality or degree of productivity, it is an indicator of the extent of use....

(817)

The hapax/type measure was developed from Aronoff’s (1976) ‘index of productivity’ (Baayen and Lieber, 1991:838). However, as mentioned earlier, Baayen and Lieber’s ‘index of productivity’ is the inverse of that proposed by Aronoff, and “expresses the extent to which the number of types in the population [...:] exceeds the number of types in the sample” (1991:838). They argue that this “provides a productivity ranking [of the affixes] that is intuitively more or less in harmony with our intuitive judgements.” (1991:838). As with the other methods examined in the previous two sections, this notion of the hapax/type ratio is particularly taken up again later, since Van Marle (1992), for instance, suggests that this ratio is, in fact, the most pertinent (158).

Baayen and Lieber’s (1991) hapax/type formula, whereby the total number of hapaxes for a particular process is divided by the total number of types with that specific affix, is expressed as:

$$\text{Productivity} = \frac{\text{Total number of hapaxes with given affix}}{\text{Total number of types with given affix}}$$

If this formula is applied to the negative prefixes, the results are as follows:
Table 4.6. Negative Prefixes – hapaxes, types and hapaxes/type figures in the BNC

<table>
<thead>
<tr>
<th>Prefix</th>
<th>No. of HAPAXES</th>
<th>No. of TYPES</th>
<th>HAPAXES/TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-</td>
<td>1,998</td>
<td>3,784</td>
<td>0.53</td>
</tr>
<tr>
<td>un-</td>
<td>1,468</td>
<td>3,836</td>
<td>0.38</td>
</tr>
<tr>
<td>a-</td>
<td>50</td>
<td>150</td>
<td>0.33</td>
</tr>
<tr>
<td>dis-</td>
<td>112</td>
<td>429</td>
<td>0.26</td>
</tr>
<tr>
<td>in-</td>
<td>189</td>
<td>922</td>
<td>0.20</td>
</tr>
<tr>
<td>All five negating prefixes</td>
<td>3,817</td>
<td>9,121</td>
<td>0.34</td>
</tr>
</tbody>
</table>

As Table 4.6 shows, hapaxes account for over half of all non- types, over a third of all un- types and exactly a third of all a- types. Dis- and in- have the lowest proportion of hapaxes per type. As with the hapax/token method, this would indicate that non- is the most productive of the negative prefixes, whilst dis- and in- are the least productive.

4.1.5.1 Discussion of the hapax/type method – Measure C

As discussed above, Baayen and Lieber’s hapax/type method proposes that a high hapax/type ratio signifies a productive affix, and that this method allows the affixes to be ranked in terms of their productivity (1991:810). This measure is premised on similar assumptions to the previous measure, but instead of calculating the proportion of hapaxes per token, it calculates the proportion per type. This means that frequency (depth) of usage is of less interest; rather, it is the number of different words which make use of an affix (breadth) that is of concern to this measurement.

Van Marle prefers the hapax/type method to the hapax/token method, and claims that token frequency is not as relevant a variable in a measure of productivity, and argues that “(o)nce a word is coined, the frequency in use of that word, it seems to me, is more or less irrelevant to the degree of productivity of that rule.” (1992:156). More broadly, Van Marle notes that Baayen’s work is concerned with
the ‘likelihood’ of a process being productive, rather than its actual ‘ability’ to be productive (1992:152). On this basis, he challenges Baayen’s view of the usefulness of hapaxes, and argues that they require further investigation. Despite this, Van Marle maintains “that the correlation between hapaxes and types is of more interest to productivity” than that between hapaxes and tokens (1992:158).

This concludes the introduction of the three main variables and the five methods used in previous calculations of morphological productivity. The five measurements of productivity examined to date are now summarised.

4.1.5.2 Overview of the Five Measurements

So far, this chapter has introduced the five key approaches to gauge the productivity of a morphological process. The first two, type-counting and token-counting, were shown to have contributed to a small number of early studies. However, it is now widely accepted that type and token frequency alone cannot be used in isolation. Some low frequency affixes may be highly productive and some high frequency affixes may have no current productivity at all. Where the type and token distributions are useful is in the calculation of mean type frequency. Though Aronoff (1976; 1980) assumed that a higher average type/token figure was indicative of high productivity, subsequent studies, such as Baayen and Lieber (1991), have assumed the opposite – that it is a low number of tokens per type that signals productivity. The latter view is based on the premise that productive affixes constantly generate new words, and thus keep the mean type frequency low. The logical extension of this, as Baayen and his colleagues noted (Baayen and Lieber, 1991; Baayen, 1992 and 1994; and Baayen and Renouf, 1996), is that the distribution of hapaxes in a corpus must have something to say about productivity. The final two measurements
therefore considered the proportion of hapaxes per token and hapaxes per type
respectively.

All five measures are returned to and critiqued more closely later in the
chapter. Before then, this chapter continues to replicate elements of previous studies
and to present further findings from the database of negatively-prefixxed words that
might be considered relevant to studies of their productivity, especially evidence that
relates to the key variable of hapax legomena. The next section revisits Baayen and
Lieber’s work on once-, twice- and thrice-occurring words in the corpus, and applies
the measure which they propose to the negating prefixes under scrutiny.

4.1.5.3 Frequencies of low-occurring words: the hapax legomena, dislegomena
and trislegomena

Baayen and Lieber claim that productive affixes exhibit a large proportion of types
which are hapaxes, which they illustrate by means of a histogram for the frequency
distribution for the suffix *-ness* (1991:810). Their histogram plots the number of
once-, twice- and thrice-occurring words in the corpus. They believe that the word-
frequency distribution for productive affixes is “highly skewed to the left”, in other
words, that the absolute number of types is greatest “at the left-hand edge” when
demonstrated by means of a histogram bar chart (1991:810). Their logic is that,
because low frequency words are indicative of the creation of new words, one would
expect a more productive affix to fall sharply, from having a high proportion of one-
token types to a much lower proportion of two-token, and then three-token, types.
According to their research, Baayen and Lieber claim that the suffix *-ness* is a
productive affix, since their histogram shows it as sharply skewed to the left
(1991:810). This method is replicated for the five negative prefixes in Figure 4.2.
The horizontal axis of Figure 4.2 below shows the frequencies in terms of types which occur only once (hapax legomena), twice (‘dislegomena’\(^6\)), and three times (‘trislegomena’\(^7\)). The vertical axis shows the proportion of types with that specific frequency. The five negative prefixes under consideration here are shown side-by-side in a histogram with their respective hapax legomena, dislegomena and trislegomena frequency distributions.

**Figure 4.2. Histogram showing hapax legomena, dislegomena and trislegomena type-frequency distributions of the negative prefixes – non-, a-, un-, dis- and in-.**

![Histogram showing frequency distributions of negative prefixes](image)

Figure 4.2 shows that the prefix which is most highly skewed to the left, that is, the one which shows the steepest incline, is *non-* , while the prefix which is least skewed to the left is *in-* . It is interesting to note that *non-* is even more skewed to the left than Baayen and Lieber showed *-ness* to be. As discussed earlier, if *-ness* has been described as a productive affix, this would indicate that *non-* is an even more productive affix.

As outlined above, Baayen and Lieber propose that productive affixes include a considerable percentage of types which are hapaxes (1991:810). The

---

\(^6\) My terminology for words which occur twice in a given corpus.

\(^7\) My terminology for words which occur three times in a given corpus.
previous section examined this phenomenon in relation to the negative prefixes, over
the frequency range of one to three. According to Baayen and Lieber’s measure, the
more productive affixes demonstrate the steepest incline to the left (1991:810). The
histograms of Figure 4.2 are also demonstrated more comparatively and pictorially in
a line chart in Figure 4.3 below.

Figure 4.3. Chart showing the negative prefixes over a type frequency range of 1-3

4.1.5.4 Analysis of type frequencies over greater frequency-distribution range
The previous section replicated Baayen and Lieber’s type frequency histogram over
the frequency range of one to three, and found that, in the case of the negative
prefixes, non- demonstrates the sharpest incline to the left, which according to this
measure, would make it the most productive of the five prefixes. To add further
support to this finding, the type-frequency distributions for the five negative prefixes
were also examined over a greater range of frequencies, in order to get a clearer
overview of the pattern emerging. To facilitate this, the database was examined in
terms of types, and the total number of types with frequencies between one and five was counted for each of the prefixes. This was then repeated for all types with frequencies between six and ten, between eleven and fifteen, and finally between sixteen and twenty.

This study, broadly-speaking, accepts Baayen and Lieber’s principle that the degree of productivity of an affix can be revealed by the type-frequency distribution of the various affixes by means of the steepness of the incline (the degree to which the distribution is skewed to the left in a histogram chart). Baayen and Lieber’s application, however, typically relates to affixes which are in most cases unrelated to each other. For instance, they investigate the suffixes -ness and -ity which, though both abstract noun-forming suffixes, are not usually in competition with each other. However, the current research examines affixes which can and often do compete for the same environment and, therefore, the ability to make comparisons in this way is a more useful and meaningful measure in the case of the negative prefixes. The counts for all the prefixes at the various frequencies are shown in Table 4.7.

| Table 4.7. Negative Prefixes – types, frequencies 1-5, 6-10, 11-15, 16-20 and frequency/type ratio (after Baayen and Lieber, 1991) |
|-------------------------------------------------|------------------|--------------|-----------------|-----------------|
| **No. of TYPES** | **Frequencies 1-5** (Proportion of all TYPES) | **Frequencies 6-10** (Proportion of all TYPES) | **Frequencies 11-15** (Proportion of all TYPES) | **Frequencies 16-20** (Proportion of all TYPES) |
| **non** | 3,784 | 3,269 (0.86) | 277 (0.07) | 107 (0.03) | 60 (0.02) |
| **un** | 3,836 | 2524 (0.66) | 299 (0.08) | 156 (0.04) | 91 (0.02) |
| **a** | 150 | 57 (0.38) | 15 (0.01) | 9 (0.06) | 3 (0.02) |
| **dis** | 429 | 203 (0.47) | 42 (0.10) | 19 (0.04) | 11 (0.03) |
| **in** | 922 | 379 (0.41) | 83 (0.09) | 38 (0.04) | 45 (0.05) |
Thus, the proportion of all types at each of the frequency ranges was calculated for each of the negative prefixes, which is demonstrated in Figure 4.4 below.

**Figure 4.4. Histogram showing 1-5, 6-10, 11-15 and 16-20 type-frequency distributions of the negative prefixes**

The horizontal axis of the chart shows the type frequencies ranges for 1 to 5, 6 to 10, 11 to 15 and 16 to 20. Figure 4.4 shows that, again, the prefix most highly skewed to the left is non-, with the prefix in-, again, demonstrated as the least skewed to the left. These same figures are more graphically illustrated in Figure 4.5, which is a chart again showing the negative prefixes over a type frequency range of 1-20.

**Figure 4.5. Chart showing 1-5, 6-10, 11-15 and 16-20 type-frequency distributions of the negative prefixes**
The notion of the type-frequency distributions of the negative prefixes is discussed in further detail in the subsequent section, which compares the 1-3 range and the 1-20 range.

4.1.5.5 Comparison of type frequencies over the 1-3 range and the 1-20 range

Figures 4.3 and 4.5 are remarkably similar. In both cases, *non-* demonstrates the steepest incline (the most highly skewed to the left), with over 50% of all *non-* types being hapaxes. According to the tables, *non-* is the most productive of the five negative prefixes. In Figure 4.3, *in-* shows the shallowest incline, is the least skewed to the left and, therefore, this makes it the least productive of the negative prefixes according to this measure. Initially, in Figure 4.5, it appears to show that the prefix *a-* has the shallowest incline, but on closer inspection this is, in fact, not the case, and it is again *in-* which is least skewed to the left. The prefix *a-* has a lower proportion of types overall in the ‘1-5’ and ‘6-10’ frequency ranges, but demonstrates a higher proportion of types than all the other negative prefixes in the ‘11-15’ type-frequency range.

Baayen and Lieber’s (1991) application of the above measures typically engages with affixes which are class-changing and in most cases are unrelated affixes (810 and 821). The prefixes examined here, however, are affixes which are class-maintaining and are those which can and often do compete for the same environment. In some respects, therefore, the above calculations are even more relevant than those assessed by Baayen and Lieber in 1991, since they provide the facility to make parallels between comparable affixes in a profitable and meaningful way.
4.1.6 A critique of Measures A, B and C

The previous five sections (Sections 4.1.1 to 4.1.5) have examined five different methods which have been used in measuring productivity. The first two methods, involve raw type and raw token data, and have been discounted, since neither type nor token frequencies individually have much to say about newly-coined word formations or innovation in language, nor, therefore, do they engage with the notion of productivity as discussed by this thesis.

Tottie, 1980; Baayen and Lieber, 1991; Baayen, 1992; Frauenfelder and Schreuder, 1992; Van Marle, 1992; Schreuder and Baayen, 1994; Baayen and Renouf, 1996; Kwon, 1997; Plag, 1999; and Plag et al, 1999 have all, to greater or lesser extents, discussed the relevance of type frequency and token frequency in terms of measuring productivity. As discussed above, in the pre-corpus era, type frequency was seen as the means of accessing information on productivity and this was often achieved via the use of dictionaries, whereas more recently, with the event of corpora, token frequency became accessible and has also been viewed as a relevant factor in assessing productivity. However, it is only when considered in combination with other variables that type and token frequencies may have worthwhile contributions to make in measures of productivity.

Measure A, the ‘type/token’ method, was the next method to be examined. The type/token method has been investigated by Baayen and Lieber, who claim that lower mean token frequency equates to higher productivity, whereas higher mean token frequency indicates lower productivity (1991:809). The type/token ratio engages with two dimensions and thus conveys a breadth and a depth of coverage perspective, dealing with both the existence of a type as well as the extent to which that type is utilised. However, when this method is applied, it compares the total
number of types with the total number of tokens, and the method provides a ‘snapshot’ at a given moment in time. In other words, it refers to the coinages that have occurred up to that point, and allows comparison with other affixes, but it does not specifically engage with the innovative use of affixes. It captures the moment in a single frame, rather than explores the wider view.

Measure B is the hapax/token measure. Baayen first introduced the notion of hapax legomena as a potential variable in 1989, and Baayen and Lieber (1991) developed the approach further in terms of a hapax/token measure (809), whereby they claim that affixes with low token frequencies are more likely to be used to create neologisms and that these lower token frequencies therefore indicate higher productivity (810). Effectively, the hapax/token ratio compares words from the corpus which occur only once with the total number of actual instances of all the types with a particular affix.

Measure C also makes use of the notion of hapaxes, this time in respect of a hapax/type method. Again, Baayen and Lieber (1991) propose that a high hapax/type ratio indicates the presence of a productive affix (810). As with Measure B, the way in which hapaxes are used for this method requires further and more sophisticated development. Once more, a measure which involves hapaxes refers to likelihood rather than fact. However, if a list of hapaxes contains neologisms, then hapaxes can allow access to these neologisms, and subsequently, these can unlock the access to the genuine innovative formations. Similarly, the notion of type refers to the number of words created by the addition of a particular affix and, therefore, in conjunction with hapaxes, these variables should reveal something about the productivity of an affix. As referred to above, however, the notion of hapax legomena requires more scrutiny.
It is impossible to come up with a definitive statement in terms of which prefix is the most productive, but a wealth of evidence now points towards extremely similar rankings, with *non-* consistently shown as the most productive of the five negative prefixes, and *dis-* and *in-* unfailingly in fourth and fifth positions respectively. Table 4.8 illustrates the rankings of Measures A, B and C.

**Table 4.8. Comparing the productivity measures (A, B and C) for the negative prefixes**

<table>
<thead>
<tr>
<th></th>
<th>Type/Token Ratio</th>
<th>Hapax/Token Ratio</th>
<th>Hapax/Type Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>non-</em> 0.168</td>
<td><em>non-</em> 0.089</td>
<td><em>non-</em> 0.53</td>
</tr>
<tr>
<td>2</td>
<td><em>a-</em> 0.045</td>
<td><em>a-</em> 0.015</td>
<td><em>un-</em> 0.38</td>
</tr>
<tr>
<td>3</td>
<td><em>un-</em> 0.022</td>
<td><em>un-</em> 0.009</td>
<td><em>a-</em> 0.33</td>
</tr>
<tr>
<td>4</td>
<td><em>dis-</em> 0.011</td>
<td><em>dis-</em> 0.003</td>
<td><em>dis-</em> 0.26</td>
</tr>
<tr>
<td>5</td>
<td><em>in-</em> 0.008</td>
<td><em>in-</em> 0.002</td>
<td><em>in-</em> 0.20</td>
</tr>
</tbody>
</table>

Table 4.8 shows that the three measures developed by Baayen (1989) and Baayen and Lieber (1991) all produce similar results, in terms of ranking at least, when applied to negating prefixes in the *BNC*. In particular, *non-* is shown to be the most productive of the five prefixes. This is consistent with the results of dictionary-based methods, for example, Hulse (2002) which is based on Bolozky’s (1999) dictionary measure, as described in the previous chapter (Section 4.1.1.2), and also with the intuitions of a number of researchers, such as Marchand (1969:180) and Mazzon (2004:111), who, though separated by some thirty-five years, both noted that the prefix *non-* was gaining in its level of productivity.

All three measures applied also agree that *dis-* and *in-* are currently the least productive of the five prefixes and that, although their quotients are very similar, *dis-* is marginally the more productive of the two. The only difference in the three sets of rankings involves *a-* and *un-* , whereby Measures A and B point towards *a-* as the more productive, whereas Measure C favours *un-* . The reason for this divergence is
that the first two measures involve token frequency as a variable but the third does not. Token frequency makes a considerable difference when comparing the two prefixes because un- occurs on 170,984 occasions in the BNC but a- occurs only 3,347 times. This means that when either types or hapaxes are divided by the number of tokens for a given prefix (as in Measures A and B respectively), a comparatively low figure is generated.

In turn, this raises the question of whether token frequency should be incorporated into the formulae for measuring productivity at all. Baayen appears to favour the hapax/token method, though gives no definitive reason for this (save that the hapax/type method is a less specific measure than the hapax/token measure). Advocates of tokens, such as Baayen, would argue that they reflect ‘depth’ of usage, and are therefore a reliable indicator of everyday language use and of the extent to which speakers of a given language are exposed to a particular affix. Critics, such as Van Marle, on the other hand, prefer the hapax/type approach, and argue that tokens are an unnecessary variable that artificially distort productivity measures, since the amount of usage becomes irrelevant once coinage has taken place (1992:156). Following Van Marle, this thesis prefers the latter position. In order to see why, consider the frequency of the top ten words in the BNC prefixed with un- and a-, as shown in Table 4.9 below:

<table>
<thead>
<tr>
<th>un-</th>
<th>Frequency</th>
<th>a-</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>unemployment</td>
<td>6,412</td>
<td>anaesthetic</td>
<td>376</td>
</tr>
<tr>
<td>unable</td>
<td>6,141</td>
<td>apathy</td>
<td>276</td>
</tr>
<tr>
<td>unlikely</td>
<td>5,547</td>
<td>asymmetry</td>
<td>251</td>
</tr>
<tr>
<td>unfortunately</td>
<td>4,553</td>
<td>amorphous</td>
<td>208</td>
</tr>
<tr>
<td>un(-)known</td>
<td>4,271</td>
<td>atheist</td>
<td>166</td>
</tr>
<tr>
<td>unusual</td>
<td>4,032</td>
<td>anaerobic</td>
<td>150</td>
</tr>
<tr>
<td>unlike</td>
<td>3,902</td>
<td>a(-)typical</td>
<td>144</td>
</tr>
<tr>
<td>unemployed</td>
<td>2,760</td>
<td>asynchronous</td>
<td>140</td>
</tr>
<tr>
<td>uncertainty</td>
<td>2,600</td>
<td>asymmetric</td>
<td>139</td>
</tr>
<tr>
<td>undoubtedly</td>
<td>2,343</td>
<td>asymptomatic</td>
<td>129</td>
</tr>
</tbody>
</table>
It is clear that *un*-words tend to be higher frequency than *a*-words. Table 4.4 showed that the mean type frequency of *un*-words (44.57) was double that of *a*-words (22.31), and Table 4.9 (above) indicates that the highest frequency *un*-words occur at much higher rates than the most common *a*-words. But are these frequency levels indicative of lower productivity, as the logic which underpins Baayen’s formulae implies? The position taken by this thesis is that, in general, token frequency is the least appropriate of the variables used in the previous studies. The rationale for this is the ubiquity of words such as *unemployment, unable* and *unlikely* in the language, which cannot be attributed to their prefix alone. If this research was replicated in the future using a newer version of the *BNC*, and the frequency of, say, *unemployment* was found to have doubled, one could not conclude that the productivity of *un*-had fallen as a result of this increased token frequency. However, this is exactly what Measures A and B would show, albeit by a negligible amount.

The rationale for calculating mean token frequency has, however, some merit. A more productive affix undoubtedly generates more new coinages more often than a less productive affix. This would normally contribute towards keeping the mean token frequency relatively low. Yet, the problem is that other factors may be equally responsible for lowering mean token frequency, for example, a prefix could be associated with words which are becoming obsolete in the language. Furthermore, as will be shown in the domain analysis presented later in this chapter (Section 4.4), it is clear that some affixes are more common in certain genres. If particular genres contribute only a small amount to total language use (that is, if they do not tend to occur in everyday language), then the frequency of established words in these areas may never reach particularly high levels, though this bears no relation to the productivity of an affix. Token-based formulae could, therefore, exaggerate the
productivity of affixes that incline towards specialist fields, compared to those that incline towards general everyday use. For example, as will be demonstrated in Section 4.4, *a*-*prefixed words are more likely to be found in the field of Natural Science; an area which is less likely to be part of our daily language, whereas words prefixed with *dis*-, on the other hand, more commonly occur in the language of World Affairs and might, therefore, be expected to be incorporated more into the language of the media and news reporting. However, whether this uneven genre distribution makes *a*- more productive than *dis*- is another matter.

The aim of the type/token measurement is to reflect new usage in a language, and a more direct way of achieving this is to make use of the hapaxes, as Baayen and his colleagues discovered. For the reasons discussed above, this research claims that dividing the number of hapaxes in a corpus by the total number of types, not tokens, provides the more reliable indicator of productivity. However, Baayen’s assumption that hapaxes are a good indicator of productivity purely because they may include a high proportion of neologisms is one that has never been fully tested. The next section, therefore, provides a closer inspection of once-occurring, negatively-prefixed words in the *BNC*. Given that hapaxes are used so prominently in the study of morphological productivity, relatively little attention has been paid to their characteristic properties. Specifically therefore, the following section examines the actual extent to which hapaxes reflect the coinage of new words.

### 4.2 Examining the hapaxes

Baayen and Renouf (1996) state that “large numbers of hapax legomena are a sure sign that an affix is productive” (74) and go on to suggest that “(n)eologisms are significantly overrepresented among hapax legomena” (76). This is the premise upon
which much of Baayen’s and his colleague’s methodology rests; that the presence of a new word or neologism in a language is inherently linked to language productivity; and that neologisms are the fruits of a productive system. As referred to in Chapter 2, Baayen (1989), Baayen and Lieber (1991) and Baayen and Renouf (1996) propose that a word which occurs only once in a given corpus is, in most cases, a newly-created word of the language, and propose that productivity can be calculated by a measure based “on the number of hapax legomena for a given affix in a corpus” (Baayen and Lieber, 1991:801). This research, however, demonstrates that most hapaxes are not, in fact, neologisms, and that the proportions which are neologisms vary considerably from affix to affix. While the first finding will have its implications for measuring productivity, the second finding has more far-reaching consequences since, if the number of neologisms in a list of hapaxes varies between different affixes, then a fundamental principle of Baayen and his colleagues is brought into question. Many single occurrence or low frequency words may indeed be newly coined words, but examination is needed to ascertain the degree to which this is the case for each affix. This section clarifies the extent to which hapaxes represent neologisms and, therefore, the extent to which hapaxes reflect productivity.

Clearly, a list of hapaxes is more than likely to include a notable proportion of words which are simply rare words of the language, words which may be becoming less popular, or words which are beginning to fall out of use all together. As noted in the literature review, Van Marle (1992) argues that Baayen’s methods rely too heavily on the value of hapaxes, and that hapaxes do not necessarily imply the presence of neologisms. He questions “whether Baayen’s approach to equate hapaxes with newly coined formations (and, consequently, to consider them indicators of productivity) is generally tenable” (1992:157). In Van Marle’s view,
once a word has been coined, the productivity has already occurred, and that the
subsequent frequency of use of that word is irrelevant (1992:157). Van Marle does
not entirely reject the usefulness of hapaxes, but proposes that more in-depth
investigation is needed, with the use of much larger corpora than was available to
Baayen in the late 1980s and early 1990s. He argues that “the notion of hapax to be
in need of further sophistication.” (1992:157). Whilst Baayen and his collaborators
(Baayen, 1989; Baayen and Lieber, 1991; Baayen and Renouf, 1996) do not
explicitly claim that all hapaxes are neologisms, they do imply that a word which
occurs only once in a given corpus is, typically, a newly-created word. The current
research supports Van Marle’s view that productivity measures should focus less on
hapaxes and more on the presence of neologisms. Therefore, it is important to
examine and investigate neologisms in order to attempt to define the notion.

The term ‘neologism’ literally means ‘new word’, yet defining a neologism is
far from simple. Matthews in The Oxford Concise Dictionary of Linguistics
describes a neologism as “Any new word which is introduced into a language, by
whatever process.” (1997:241). Certainly, a neologism falls somewhere between a
nonce word (a word made up on the spur of the moment to fulfil an immediate need)
and an established word of the language, though it is not easy to decide the exact
point at which a new word is introduced into the language, and the notion remains
ambiguous and disputed by linguists. Bauer proposes that “a neologism is a word
which becomes part of the norm of the language, and thus is part of the brief of a
lexicographer” but also goes on to state that “(i)t is probably not possible to tell at
the point when a word is coined whether it will turn out to be a nonce word or a
neologism” (2001:39). Thus, some contend (Crystal, 1997a) that a word becomes a

8 My italics.
neologism at the point when a nonce word is used by someone other than the creator. For others (Bauer and Lieber, 1991), it is when a new lexical item begins to gain popularity and more widespread usage, but is, as yet, not included in a dictionary. For others still (Bauer, 2001), it is the point when a word is included in a dictionary for the first time. Equally unclear, ‘included in a dictionary’ could be taken to mean any dictionary; a specific dictionary; or a recognised standard in terms of dictionaries, such as the Oxford English Dictionary (OED).

For Baayen and Lieber, a neologism is a lexical item not yet established in the language, and “not registered in the most comprehensive standard dictionary” (1991:812-813). However, if the notion of neologism is imprecise, and if hapaxes are simply low frequency words, rather than neologisms, this raises several important issues in respect of their methods for measuring productivity. The following section, therefore, clarifies the connection between hapaxes and neologisms, whereby a sample of negatively-prefixed hapaxes are selected from the database.

Commenting on Baayen’s view, Plag (2003) points out that, as hapaxes have the lowest possible frequencies, there should be some relationship between them and the number of neologisms. However, it is only in this sense that hapaxes can be seen as a reflection of productivity. Plag proposes, therefore, that:

…the number of hapaxes of a given morphological category should correlate with the number of neologisms of that category, so that the number of hapaxes can be seen as an indicator of productivity.

(2003:54)

Neither Plag (2003) nor this research propose that hapaxes and neologisms are equivalent, and certainly there is no absolute and consistent relationship between the two phenomena. Lieber maintains that:
The more productive a process is, the more new words it will give rise to and the more chance that these items will occur in a corpus with a very low frequency, sometimes only once. [...] The ratio of hapaxes to tokens tells us something about productivity.

(2010:67)

Lieber also acknowledges that Baayen’s original (1989) measure is a “...sophisticated – but still not perfect – measure of productivity” (2010:67). Nevertheless, Baayen’s measures still stand as innovative and ground-breaking at the time of their introduction. Bauer, however, proposes that “(w)e need to ask why hapaxes in a corpus should correspond in any meaningful way to coinages in real use” (2001:150). By their nature, most errors in a corpus end up in a list of hapaxes, since they are typically likely to be once-occurring words, though clearly these have nothing to do with productivity. The examples disbeliever and unfortunately, for instance, do not comment on the productivity of the prefixes dis- and un- in any way. This fact alone demonstrates that hapaxes need further investigation. Likewise, though the example unbuttonable is a hapax and is not included in the OED, this may not, again, reflect on the productivity of un- so much as the suffix -able.

Clearly, a word may be low frequency for a variety of reasons, and it is important to establish the extent to which individual hapaxes can be regarded as neologisms. Indeed, this is one of the key issues of this research. Although a translation by Van Marle who criticises Baayen on a number of issues, Baayen is reported as stating that “(t)he larger the number of hapaxes [...], the higher is the degree of productivity” (Baayen, 1990:218, cited in Van Marle, 1992:156). Whilst Van Marle does not bring into question Baayen’s view of the relationship between frequency and neologisms, he argues that “this relationship is much less transparent than Baayen wants us to believe” (1992:157), and that Baayen’s concept of
productivity results from the “doubtful status assigned to hapaxes” (1992:161).

Further, Van Marle goes on to argue that:

Baayen is not so much concerned with the *ability* of morphological processes to underlie newly coined words, but [...] *with the degree to which the morphological processes involved may actually be expected to give rise to newly coined words.*

(1992:152)

In other words, despite the corpus-based approach, Baayen’s hapax-related measures rely on probability rather than on fact, and if hapaxes are simply low frequency words, rather than neologisms, then this may challenge Baayen’s methods of measuring productivity.

### 4.2.1 Identifying the correlation between hapaxes and neologisms

A sample of hapaxes was extracted from the database: approximately 20% of the total number of hapaxes (3,817). Specifically, a random sample of 200 hapaxes was taken for each of the negative prefixes in the database, unless there were fewer than 200 hapaxes for that particular prefix, in which case, all the hapaxes for that prefix were sampled. A total number of 751 hapaxes were, therefore, examined. This comprised all 50 *a*-prefixed hapaxes, all 112 *dis*-prefixed hapaxes, all 189 *in*-prefixed hapaxes, 200 *non*-prefixed hapaxes and 200 *un*-prefixed hapaxes. To ensure the consistency of the application of the methodology, criteria were determined in order to verify whether a specific hapax was indeed a neologism or an established word of the language, and all of the sampled hapaxes were assigned to a particular category based on this classification.

---

9 Van Marle’s italicising.
Firstly, a small number of sample hapaxes were found to be incorrectly spelt words (e.g. *imprevisable* for *imprevisible; unconformably* for *unconformably*). These were discounted, since correctly spelt, they would not constitute hapaxes. Secondly, in cases where the sample hapax was made up of three or more components (i.e. a negative prefix, and a base, in addition to one or more suffixes), the form was checked for inclusion in the *OED*. If the form of the word which incorporated the prefix and the base word was included in the *OED*, but the complete complex word incorporating the suffix was not listed, this was classified as a ‘neologism via suffixation only’. For instance, the hapaxes *discreditability, indecipherably* and *unbuttonable* do not appear in the *OED*, even though *discredit* (1559), *indecipherable* (1802) and *unbutton* (c.1325) do. Therefore, *discreditability, indecipherably* and *unbuttonable* are neologisms in terms of their suffixes only. Thirdly, each of the 751 sample hapaxes was manually examined for its inclusion in the *OED*. If the sample hapax was included in the *OED*, in exactly the same form, its date of entry into the language was recorded, it was discounted as a neologism, and it was classified as an established word of English. All of the words which did not fall into any of the above categories were classified as genuine neologisms. Whilst the *OED* may have certain limitations in terms of its age and the possibility that semantically transparent items may not be included, nevertheless it remains the most comprehensive dictionary of its kind. As Plag suggests:

Students of English [....] are in the advantageous position that there is a dictionary like the *Oxford English Dictionary* (*OED*). This dictionary has about 500,000 entries and aims at giving thorough and complete information on all words of the language and thus the development of the English vocabulary from its earliest attestations onwards.

(2003:52)

---

10 Bracketed items refer to date of entry into the language.
The choice of the *OED* as a single source can be justified in that it provides a variety of information that other dictionaries do not. In addition to the usual information provided by dictionaries, the *OED* also contributes information on etymology, quotations and dates of these quotations, earliest attestations, etc, which “allows us to trace neologisms for any given period in time (Plag, 2003:75).

### 4.2.2 The application of disqualification criteria to determine the ‘genuine’ neologisms

Section 4.2.1 above establishes a number of criteria, and provides a rationale, for not regarding all hapaxes as neologisms which, in turn, enabled a classification of the hapaxes in the *BNC/database*. These disqualification criteria, therefore, are as follows:

(i) Misspellings and other errors were discounted.

(ii) Neologisms as a result of suffixation only were discounted.

(iii) Words listed in the *OED* were classified as established words and were discounted.

These disqualification criteria were then applied to the 751 sample hapaxes in order to classify their status and to establish the genuine neologisms in the sample. The results are shown in Table 4.10 below. The findings presented in the table have significant ramifications for the methods used in previous corpus-based studies of morphological productivity. Firstly, there were some negatively-prefixed sample hapaxes in the database which were only hapaxes in terms of error (for example, misspellings): approximately 3% over all the prefixes.
Table 4.10. Application of disqualification criteria to filter ‘genuine’ neologisms

<table>
<thead>
<tr>
<th></th>
<th>a-</th>
<th>dis-</th>
<th>in-</th>
<th>non-</th>
<th>un-</th>
<th>Overall percentages per category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misspellings and other errors</td>
<td>0.0%</td>
<td>4.0%</td>
<td>7.0%</td>
<td>1.0%</td>
<td>1.0%</td>
<td>3.0% (21)</td>
</tr>
<tr>
<td>Neologism via suffixation only</td>
<td>2.0%</td>
<td>13.0%</td>
<td>15.0%</td>
<td>1.5%</td>
<td>15.0%</td>
<td>10.0% (77)</td>
</tr>
<tr>
<td>Established word included in OED</td>
<td>78.0%</td>
<td>71.0%</td>
<td>72.0%</td>
<td>12.5%</td>
<td>60.5%</td>
<td>53.0% (401)</td>
</tr>
<tr>
<td>Non-neologism totals</td>
<td>80.0%</td>
<td>88.0%</td>
<td>94.0%</td>
<td>15.0%</td>
<td>76.5%</td>
<td>66.0%</td>
</tr>
</tbody>
</table>

‘GENUINE’ NEOLOGISM

<table>
<thead>
<tr>
<th></th>
<th>20.0%</th>
<th>12.0%</th>
<th>6.0%</th>
<th>85.0%</th>
<th>23.5%</th>
<th>34.0% (252)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of hapaxes sampled</td>
<td>(50)</td>
<td>(112)</td>
<td>(189)</td>
<td>(200)</td>
<td>(200)</td>
<td>100.0% (751)</td>
</tr>
</tbody>
</table>

(Figures rounded to the nearest .5 or .0)

Secondly, the table demonstrates that, overall, the hapaxes comprise approximately 10% of established negatively-prefixed words which have been further derived by means of the addition of suffixes and are, therefore, only neologisms by virtue of their suffixation. These neologisms via suffixation can be discounted as they do not reflect the productivity of negative prefixes. Thirdly, with regard to established words, the table reveals that over 50% of all the hapaxes in the sample are already included in the dictionary. Indeed, in the case of the prefixes a-, dis- and in-, this proportion is actually over 70%. The high percentage of established words in the hapax sample is found across all the prefixes except non-, though even the percentage for non- is not inconsequential (12.5%).

This leaves the negatively-prefixed hapaxes which are deemed to be neologisms. Over all the prefixes, the percentage of hapaxes which may be classed as neologisms is 34%. However, this average is somewhat misleading, since the percentages vary considerably from prefix to prefix. Only 6.0% of in- prefixed hapaxes fall into this category of genuine neologisms, with 12.0% of dis- prefixed hapaxes; 20.0% of a- prefixed hapaxes; and 23.5% of un- prefixed hapaxes that
could be classified as neologisms. Alternatively, however, 85% of non- prefixed hapaxes are neologisms. As the next section discusses, this is of major significance to the formulae used by researchers such as Baayen (1989) and his collaborators, whose comparisons are implicitly based on a consistent proportion of hapaxes being neologisms across all affixes.

4.2.3 To what extent do hapaxes reflect morphological productivity?

Table 4.10 shows the classification of the hapaxes, the extent to which hapaxes represent neologisms and, thus, reflect productivity. Taking the negative prefixes individually, only 6.0% of the in- prefixed hapax sample are neologisms, compared to 85.0% of non- prefixed hapaxes which are neologisms. This demonstrates that there is a distinctly uneven distribution from prefix to prefix, and this makes direct comparisons between the different negative prefixes (or indeed comparisons between any other affixes) statistically inappropriate. In Baayen’s measures, this uneven distribution is never considered a possibility and, as a result, is never taken into account.

Based on the findings from the sampling, hapaxes cannot be considered neologisms per se, and do not straightforwardly, therefore, reflect productivity. Any list of hapaxes will, however, include a certain number of neologisms and, therefore, any list of this kind can be examined in greater detail in order to access the incorporated neologisms. However, without quantifying the neologisms, this method relies on likelihood and potential rather than on actuality. Hapaxes are likely to include a considerable number of words which are rarely-used but long-established words of the language. Hapaxes are therefore accepted by this research as a useful
vehicle to help quantify the neologisms, but they require the application of more fine-grained measures to bring out their true profitability.

According to Table 4.10, 85% of non- hapaxes are, in fact, neologisms, although this is in sharp contrast to in- hapaxes, of which only 6% are neologisms. Thus, 94% of in- hapaxes are actually not neologisms, with over 70% of these being established words of the language. Indeed, across all the negatively-prefixed hapaxes, over 50% are established words. Over three-quarters of dis- (88%), a- (80%) and un- (76.5%) hapaxes are not neologisms. Of these non-neologism percentages, as well as the established words, in- (15%), un- (15%) and dis- (13%) also demonstrate notable proportions of neologisms via suffixation, which suggests that these prefixes are more likely to be examples of further-derived words. Hapaxes prefixed with a- and non-, on the other hand, appear not to have a tendency towards further derivation. Indeed, in the case of dis- and in- hapaxes, there is a greater likelihood of encountering further-derived words than there is of finding neologisms. This serves to support the view that the prefixes dis- and in- are currently unproductive prefixes, but typically occur in more established words of the language which have come to incorporate further derivations. For example, hapaxes such as discreditability and indiscriminateness reveal nothing about the productivity of their prefixes. This also indicates how some of the productivity quotients calculated by Baayen and his colleagues have been distorted, since the proportion of hapaxes which are neologisms varies so considerably from prefix to prefix.

What this research has verified is that a variable proportion of hapaxes are neologisms. It has, similarly, determined that hapax legomena are indeed relevant in measuring productivity, as proposed by Baayen (1989), Baayen and Lieber (1991) and Baayen and Renouf (1996), though it has also been shown that the role of
hapaxes cannot simply be taken at face value. Likewise, though a database of hapaxes will certainly reveal neologisms, any list of hapaxes, unless carefully filtered, will inevitably include a considerable proportion of words which are not neologisms. Indeed, this research has demonstrated that in this database of hapaxes only approximately one third are in fact neologisms, whereas over half the words in the database of hapaxes are actually established but uncommon words of the language. It may be predicted, therefore, that other such lists of hapaxes could well demonstrate similar proportions.

4.2.4 Proposed new methods: the ‘neologism/type’ method and the ‘neologism/token’ method

This research has shown that hapaxes are often used in calculations of morphological productivity on the basis that they have a greater than average chance of being neologisms. It has also shown, unlike previous studies which have taken hapaxes at face value, that the proportion of hapaxes that are genuine neologisms is, in most cases, relatively low and, most importantly, that this proportion differs from prefix to prefix. This is a significant finding since previous research has never considered this a possibility, let alone investigated the phenomenon further. As a result of this finding, the current research is now in a position to introduce two new measures: neologisms/types and neologism/tokens. The statistics gathered from Table 4.10 allow a ‘neologism’ variable to be applied to productivity formulae which, in turn, allows for more sophisticated measures to be generated. In order to calculate productivity by these measures, the percentage of hapaxes which are genuine neologisms are first taken. This is then multiplied by the total number of hapaxes for that prefix. The resulting figure is then divided by the total number of types or tokens (depending on the measure being applied) for that prefix in the whole
database. The neologism/token measure (Measure B1) is shown in Table 4.11, and the neologism/type measure (Measure C1) in Table 4.12.

**Table 4.11. Measure B1 - productivity via a neologism/token ratio**

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Total No. of hapaxes</th>
<th>Proportion of hapaxes that are genuine neologisms</th>
<th>Total No. of tokens</th>
<th>Adjusted Level of productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-</td>
<td>1,998</td>
<td>0.85</td>
<td>22,466</td>
<td>0.0756</td>
</tr>
<tr>
<td>a-</td>
<td>50</td>
<td>0.20</td>
<td>3,347</td>
<td>0.0030</td>
</tr>
<tr>
<td>un-</td>
<td>1,468</td>
<td>0.23</td>
<td>170,984</td>
<td>0.0020</td>
</tr>
<tr>
<td>dis-</td>
<td>112</td>
<td>0.12</td>
<td>38,368</td>
<td>0.0004</td>
</tr>
<tr>
<td>in-</td>
<td>189</td>
<td>0.06</td>
<td>108,791</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**Table 4.12. Measure C1 - productivity via a neologism/type ratio**

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Total No. of hapaxes</th>
<th>Proportion of hapaxes that are genuine neologism</th>
<th>Total No. of types</th>
<th>Adjusted Level of productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-</td>
<td>1,998</td>
<td>0.85</td>
<td>3,784</td>
<td>0.4488</td>
</tr>
<tr>
<td>un-</td>
<td>1,468</td>
<td>0.23</td>
<td>3,836</td>
<td>0.0880</td>
</tr>
<tr>
<td>a-</td>
<td>50</td>
<td>0.20</td>
<td>150</td>
<td>0.0666</td>
</tr>
<tr>
<td>dis-</td>
<td>112</td>
<td>0.12</td>
<td>429</td>
<td>0.0313</td>
</tr>
<tr>
<td>in-</td>
<td>189</td>
<td>0.06</td>
<td>922</td>
<td>0.0123</td>
</tr>
</tbody>
</table>

Measures B1 (neologism/token measure) and C1 (neologism/type measure) are remarkably similar, with both measures placing *non-* as the most productive prefix, and *in-* as the least productive. The two measures only differ in terms of their ranking of the prefixes *a-* and *un-*, which interchange with each other for the second and third places. Measure B (hapax/token) and Measure B1 (neologism/token) are comparable and have the *a-* prefix in second place in the ranking, whereas Measure A (type/token), Measure C (hapax/type) and Measure C1 (neologism/type) place *un-* in second position in the ranking.

For the reasons discussed in Sections 4.1.1.3, 4.1.5.4 and 4.1.6, this thesis argues that type frequency provides a more relevant and useful denominator than token frequency. The proposal is therefore that Measure C1, which divides
neologism frequency by type frequency, expresses morphological productivity in the most methodologically robust way of any measure considered so far. In the continuing attempt to propose a more fine-grained approach to the measurement of morphological productivity, I now turn to another key issue which has previously been overlooked; that of the role of word class.

4.3 The role of word class in morphological productivity

So far in this chapter, types and tokens have been examined separately, along with types and tokens as a ratio, and hapax/type and hapax/token ratios. In addition, hapaxes have been investigated in terms of their status as neologisms, and it has been determined that hapaxes are not necessarily neologisms, but that a list of hapaxes can be expected to contain a proportion of neologisms. Proposals have also been made with regard to the development of more sophisticated methods for measuring productivity, for example, the ‘neologism/token’ method and the ‘neologism/type’ method. The current section now turns to the examination of the role of word class and how it may affect morphological productivity, and undertakes an analysis of the negatively-prefixed words in the database and considers their status in terms of word class distribution. Many of the factors discussed in the previous sections also affect and are affected by the word class to which a negatively-prefixed word belongs. In respect of word class, this approach is also needed since the negative prefixes do not necessarily operate on a level playing field. One aim of the current section, therefore, is to discover whether each negative prefix has a predisposition towards combining with words from certain word classes, or indeed whether the individual prefixes have a tendency to shun certain classes. This section discusses these preferences from the perspective of word class and how word class may affect productivity, and the
overall aim is to calculate a productivity figure for each negating prefix according to word class. The goal is to discover whether the three measures for quantifying morphological productivity can be applied in a more grammatically-refined manner in order to shed light on the productivity levels of individual word classes and their associated prefixes, for example, to show how productive, say, the prefix *non-* is amongst adjectives. The following sections apply the mean type frequency measure (Measure A), along with the newly-proposed methods (Section 4.2.4): the neologism/token (Measure B1) and the neologism/type (Measure C1) measures, with a view to replicating these but in terms of word class productivity.

Although Baayen and Lieber (1991) did not examine productivity from the point of view of word class, this section provides such data, and generates the figures necessary to calculate, as they did, three key formulae - type/token; hapax/type; hapax/token - to each of the four open word classes. The initial section below, therefore, examines type distributions from two different perspectives: first, from the perspective of the word class (to determine whether the prefixes have any preferences towards one word class over another); second, from the perspective of the individual prefixes (to determine whether words which belong to particular word classes are more likely to choose one prefix over another). The next section then briefly compares the distribution of word classes among negatively prefixed words in the *BNC* with the distribution of word classes across the whole corpus. This information is all based on type frequency. A parallel analysis and set of tables are then presented for token frequency.
4.3.1 WORD CLASS and TYPE distribution

This section examines the distribution of the negatively-prefixed words of the database in terms of type frequency, to determine if the individual negative prefixes prefer to be combined with words from particular word classes. Similarly, the section determines whether word class types have a preference for being combined with certain prefixes. The distribution of the negatively-prefixed types of the database is then compared to the word class distribution of the BNC as a whole.

4.3.1.1 Do negatively prefixes TYPES favour adjectives, adverbs, nouns or verbs?

The first stage of the analysis was to determine the total number of types for each prefix in each word class. These figures are shown in Table 4.13 below. In the few examples where types belong to more than one word class, they have been treated separately. For example, *agnostic* occurs as an adjective and also as a noun; *dishonour* appears as a noun and as a verb. Both words were treated as two separate types for the purpose of this analysis.

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Prepositions</th>
<th>Verbs</th>
<th>Total No. of TYPES</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>a</em>-</td>
<td>88</td>
<td>16</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td><em>dis</em>-</td>
<td>143</td>
<td>27</td>
<td>166</td>
<td>0</td>
<td>93</td>
<td>429</td>
</tr>
<tr>
<td><em>in</em>-</td>
<td>351</td>
<td>193</td>
<td>358</td>
<td>0</td>
<td>20</td>
<td>922</td>
</tr>
<tr>
<td><em>non-</em></td>
<td>2,525</td>
<td>70</td>
<td>1,188</td>
<td>0</td>
<td>1</td>
<td>3,784</td>
</tr>
<tr>
<td><em>un-</em></td>
<td>2,879</td>
<td>338</td>
<td>404</td>
<td>1</td>
<td>214</td>
<td>3,836</td>
</tr>
<tr>
<td>Total types for each word class</td>
<td>5,986</td>
<td>644</td>
<td>2,162</td>
<td>1</td>
<td>328</td>
<td>9,121</td>
</tr>
</tbody>
</table>

It will be noted that for *un-*-, there is a single type of preposition, in the form of *unlike*. *Unlike* can occur as a preposition, as for instance in the contextual
example: “Unlike the other men, Peter did not suggest they meet again”, or may also occur as an adjective, as for example in: “Most of the American east coast is unlike anything found in Europe” (BNC). Unlike is, therefore, incorporated into the database as an adjectival type, but also separately listed as a prepositional type. Although unlike, as a preposition, has a token frequency of 3,902, this study confines itself to the examination of the major open word classes of adjectives, adverbs, nouns and verbs, since these are the classes which typically lend themselves to new coinages. Prepositions and other closed word classes, on the other hand, rarely produce new types. As unlike is unique as a negatively-prefixed preposition, although it will be referred to in the figures and tables for the purpose of the calculations, it will not be included in the analysis of this study.

In order to determine which negatively-prefixed types favour which word class overall, the percentage distribution of the prefixed types for each word class (adverb, adjective, noun, preposition and verb) was examined in the case of each prefix, and this distribution is shown in Table 4.14 below.

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Prepositions</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>58.67%</td>
<td>10.66%</td>
<td>30.67%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>dis-</td>
<td>33.33%</td>
<td>6.29%</td>
<td>38.70%</td>
<td>0%</td>
<td>21.68%</td>
</tr>
<tr>
<td>in-</td>
<td>38.07%</td>
<td>20.93%</td>
<td>38.83%</td>
<td>0%</td>
<td>2.17%</td>
</tr>
<tr>
<td>non-</td>
<td>66.73%</td>
<td>1.85%</td>
<td>31.39%</td>
<td>0%</td>
<td>0.03%</td>
</tr>
<tr>
<td>un-</td>
<td>75.05%</td>
<td>8.81%</td>
<td>10.53%</td>
<td>0.03%</td>
<td>5.58%</td>
</tr>
<tr>
<td>All prefixes</td>
<td>65.63%</td>
<td>7.06%</td>
<td>23.70%</td>
<td>0.01%</td>
<td>3.60%</td>
</tr>
</tbody>
</table>

Table 4.14 demonstrates that negatively-prefixed words are, by a considerable margin, more likely to be adjectives than any other word class (65.63%), with nouns making up the next highest percentage of word class types (23.70%). Adverbs (7.06%) and verbs (3.60%) are the least common word class
types respectively. Within individual prefixes, some notable patterns emerge. For example, only *dis-* appears to accommodate verbs to any consequential degree, although *un-* and *in-* verbal types are demonstrated to a notably lesser extent. *Non-* is least likely to occur as part of an adverb, whereas *in-* is the most likely prefix to be found in an adverb. *Un-* appears to avoid nouns more than any other prefix. These preferences are also demonstrated in a column chart (Figure 4.6), which shows the percentage distribution of the word classes in terms of types for each of the negative prefixes.

**Figure 4.6. Word class percentages for negatively-prefixed TYPES**

![Column chart showing word class percentages for negatively-prefixed types](image-url)

- **Verbs**
  - **a-**: 30.67%
  - **dis-**: 6.29%
  - **in-**: 38.83%
  - **non-**: 66.73%
  - **un-**: 75.05%

- **Adverbs**
  - **a-**: 10.66%
  - **dis-**: 22%
  - **in-**: 3.93%
  - **non-**: 31.39%
  - **un-**: 8.81%

- **Nouns**
  - **a-**: 58.67%
  - **dis-**: 3.33%
  - **in-**: 58.07%
  - **non-**: 10.53%
  - **un-**: 0%

- **Adjectives**
  - **a-**: 0%
  - **dis-**: 0%
  - **in-**: 0%
  - **non-**: 0%
  - **un-**: 0%
4.3.1.2 Do TYPES belonging to each word class favour different prefixes?

The previous section examined the breakdown of negatively-prefixed types from the perspective of the word classes. This section shows the distribution of the negatively-prefixed types according to word class. These percentages are recorded in Table 4.15 below, which shows how likely types belonging to each individual word class are to incorporate the negative prefixes *a-* , *dis-* , *in-* , *non-* or *un-* . As the rightmost column shows, if a word includes any one of these five prefixes, it is most likely to be *un-* but least likely to be *a-* . This is true of all of the individual classes, except for nouns, where *non-* is the most common prefix. Amongst other word classes, other preferences can also be observed. For example, verbs never incorporate the negative *a-* prefix, and very rarely take *non-*. Adjectives have a tendency to avoid *in-* , *dis-* and particularly the *a-* prefix, and likewise, nouns which incorporate *a-* are uncommon. Adverbs, too, rarely include *dis-* and particularly appear to avoid the *a-* prefix.

Table 4.15. Distribution of negatively-prefixed TYPES by word class

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Nouns</th>
<th>Adverbs</th>
<th>Verbs</th>
<th>All word classes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>a-</em></td>
<td>1.47% (88)</td>
<td>2.13% (46)</td>
<td>2.48% (16)</td>
<td>0% (0)</td>
<td>1.65% (150)</td>
</tr>
<tr>
<td><em>dis-</em></td>
<td>2.39% (143)</td>
<td>7.68% (166)</td>
<td>4.19% (27)</td>
<td>28.35% (93)</td>
<td>4.70% (429)</td>
</tr>
<tr>
<td><em>in-</em></td>
<td>5.86% (351)</td>
<td>16.56% (358)</td>
<td>29.97% (193)</td>
<td>6.10% (20)</td>
<td>10.11% (922)</td>
</tr>
<tr>
<td><em>non-</em></td>
<td>42.18% (2,525)</td>
<td>54.95% (1,188)</td>
<td>10.87% (70)</td>
<td>0.31% (1)</td>
<td>41.49% (3,784)</td>
</tr>
<tr>
<td><em>un-</em></td>
<td>48.10% (2,879)</td>
<td>18.68% (404)</td>
<td>52.49% (338)</td>
<td>65.24% (214)</td>
<td>42.05% (3,835)</td>
</tr>
</tbody>
</table>

4.3.1.3 Negatively-prefixed TYPES compared with the whole BNC according to word class

The previous section presented the word class distribution of negatively-prefixed types. The current section compares these with the word class distribution (adjective,
adverb, noun and verb) for the whole of the BNC. The purpose of this is to given a sense of how the above word class distributions for negatively-prefixed types compare to those in the English Language in general.

The total number of types for the four word classes for the whole BNC was calculated by means of corpus interface software: BNCweb (CQP Edition)\(^\text{11}\) (http://es-corp.uzh.ch/). The search engine produced frequency lists for each of the four word classes, along with total type frequencies for each word class. These figures, along with the word class distributions for the negatively-prefixed types are shown in Table 4.16 below.

### Table 4.16. Word class distribution percentages: negatively-prefixed TYPES vs. whole BNC

<table>
<thead>
<tr>
<th></th>
<th>Total types(^\text{12}) for whole BNC (839,772)</th>
<th>Total types for all negative prefixes (9120)</th>
<th>Total a-types (150)</th>
<th>Total dis-types (429)</th>
<th>Total in-types (922)</th>
<th>Total non-types (3,784)</th>
<th>Total un-types (3,836)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives</td>
<td>24.93% (209,353)</td>
<td>65.64% (5,986)</td>
<td>58.67% (88)</td>
<td>33.33% (143)</td>
<td>38.07% (351)</td>
<td>66.73% (2,525)</td>
<td>75.05% (2,879)</td>
</tr>
<tr>
<td>Nouns</td>
<td>62.42% (524,240)</td>
<td>23.71% (2,162)</td>
<td>30.67% (46)</td>
<td>38.70% (166)</td>
<td>38.83% (358)</td>
<td>31.39% (1,188)</td>
<td>10.53% (404)</td>
</tr>
<tr>
<td>Adverbs</td>
<td>1.12% (9,380)</td>
<td>7.06% (644)</td>
<td>10.66% (16)</td>
<td>6.29% (27)</td>
<td>20.93% (193)</td>
<td>1.85% (70)</td>
<td>8.81% (338)</td>
</tr>
<tr>
<td>Verbs</td>
<td>11.53% (96,799)</td>
<td>3.59% (328)</td>
<td>0.00% (0)</td>
<td>21.68% (93)</td>
<td>2.17% (20)</td>
<td>0.03% (1)</td>
<td>5.58% (214)</td>
</tr>
</tbody>
</table>

Columns 2 and 3 of Table 4.16 show that although most English words in the BNC are nouns (62.42%), the dominant class among negated words is adjectives (65.64%). Adverbs are more likely to take a negative prefix (7.06%) than their overall representation in the BNC would suggest (1.12%); and verbs are less likely to be negatively-prefixed words (3.59%; compared with 11.53% for the whole BNC).

\(^{11}\) Accessed 14/10/08.

\(^{12}\) Total adjecitval, nominal, adverbial and verbal types.

\(^{13}\) Excluding the prepositional type unlike.
The implication of these findings is discussed in greater depth in subsequent sections.

4.3.2 WORD CLASS and TOKEN distribution

The previous sections examined the correlation between word class and negating prefixes according to type frequency. The current and subsequent sections parallel the structure of these previous sections, but instead examine token frequency. In this section, therefore, the distribution of the negatively-prefixed words of the database is explored with regard to their token frequency. Again, as with type frequency, the aim is to determine whether the individual negative prefixes prefer being incorporated into words from particular word classes or if the prefixes tend to evade attachment to certain word-class bases. The section also examines whether word class tokens prefer or avoid certain prefixes. The distribution of the negatively-prefixed tokens of the database is then compared to the word class token distribution for the whole BNC.

4.3.2.1 Do negatively prefixed TOKENS favour adjectives, adverbs, nouns or verbs?

Tokens are viewed as a relevant variable in measures of productivity by a number of researchers (such as Baayen and Lieber, 1991; Baayen, 1992; Baayen and Renouf, 1996; Kwon, 1997; Plag, 1999 and Bolozky, 1999), since tokens reflect the actual usage of types. Table 4.17 shows the distribution of all the negatively-prefixed tokens in the database into their relevant word classes.
Table 4.17. Word class breakdown for negatively-prefixed TOKENS

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Prepositions</th>
<th>Verbs</th>
<th>Total No. of TOKENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>1,817</td>
<td>70</td>
<td>1,460</td>
<td>0</td>
<td>0</td>
<td>3,347</td>
</tr>
<tr>
<td>dis-</td>
<td>6,225</td>
<td>661</td>
<td>18,241</td>
<td>0</td>
<td>13,241</td>
<td>38,368</td>
</tr>
<tr>
<td>in-</td>
<td>70,079</td>
<td>15,580</td>
<td>22,372</td>
<td>0</td>
<td>760</td>
<td>108,791</td>
</tr>
<tr>
<td>non-</td>
<td>16,287</td>
<td>292</td>
<td>5,884</td>
<td>0</td>
<td>3</td>
<td>22,466</td>
</tr>
<tr>
<td>un-</td>
<td>125,210</td>
<td>18,340</td>
<td>14,231</td>
<td>3,902</td>
<td>9,301</td>
<td>170,984</td>
</tr>
<tr>
<td>All prefixes</td>
<td>219,618</td>
<td>34,943</td>
<td>62,188</td>
<td>3,902</td>
<td>23,305</td>
<td>343,956</td>
</tr>
</tbody>
</table>

It will again be noted that negatively-prefixed prepositions have a token frequency of 3,902, though this comprises the single type, *unlike*. As referred to earlier, prepositions and other closed word classes do not readily lend themselves to the productive use of new formations and, although the token frequency for *unlike* is included in the table for the purposes of the calculations, it is not discussed any further in this current research. To help determine which negatively-prefixed tokens prefer which word class, the breakdown of the token percentages for each prefix for each word class is shown below in Table 4.18.

Table 4.18. Distribution for negatively-prefixed TOKENS

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Prepositions</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>54.29%</td>
<td>2.09%</td>
<td>43.62%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>dis-</td>
<td>16.23%</td>
<td>1.72%</td>
<td>47.54%</td>
<td>0%</td>
<td>34.51%</td>
</tr>
<tr>
<td>in-</td>
<td>64.42%</td>
<td>14.32%</td>
<td>20.56%</td>
<td>0%</td>
<td>0.70%</td>
</tr>
<tr>
<td>non-</td>
<td>72.50%</td>
<td>1.30%</td>
<td>26.19%</td>
<td>0%</td>
<td>0.01%</td>
</tr>
<tr>
<td>un-</td>
<td>73.23%</td>
<td>10.73%</td>
<td>8.32%</td>
<td>2.28%</td>
<td>5.44%</td>
</tr>
<tr>
<td>All prefixes</td>
<td>63.85%</td>
<td>10.16%</td>
<td>18.08%</td>
<td>1.13%</td>
<td>6.78%</td>
</tr>
</tbody>
</table>
As was found for types, negatively-prefixed tokens are more often adjectives than any other word class (63.85%), and again, nouns make up the next highest percentage of word class tokens (18.08%). As was found with negatively-prefixed types, patterns of distribution also emerge in respect of negatively-prefixed tokens. Again, only the prefix *dis-* appears to accommodate verbs, although *un-* verbs do have some representation. *Non-* is again least likely to occur in an adverb, though *dis-* is similarly rare in adverbs, and again, *in-* is the most likely to be found in an adverb. Once more, *un-* appears to shun nouns more than any other prefix. In terms of token frequency, *dis-* is least likely to be found in adjectives. The figures shown in Table 4.18 are also demonstrated in a column chart below (Figure 4.7). The distribution of word class token frequencies is discussed further subsequent sections.

**Figure 4.7. Word class percentages for negatively-prefixed TOKENS**

![Bar chart showing word class percentages for negatively-prefixed tokens](chart.png)
4.3.2.2 Do TOKENS belonging to each word class favour different prefixes?

The previous section examined the breakdown of negatively-prefixed tokens from the perspective of the word classes. The aim of the current section is to question whether tokens which belong to each of the four major word classes show preferences for different prefixes. These figures are later used to assess the productivity levels for each of the prefixes within each word class. As with the word-class type counts, token counts help determine which negative prefix prefers which word classes and whether certain prefixes tend to avoid certain classes. The figures that relate to word-class token percentages and their distribution across the prefixes are shown in Table 4.19.

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Nouns</th>
<th>Adverbs</th>
<th>Verbs</th>
<th>All word classes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a-</strong></td>
<td>0.83%</td>
<td>2.35%</td>
<td>0.20%</td>
<td>0%</td>
<td>0.99%</td>
</tr>
<tr>
<td></td>
<td>(1,817)</td>
<td>(1,460)</td>
<td>(70)</td>
<td>(0)</td>
<td>(3,347)</td>
</tr>
<tr>
<td><strong>dis-</strong></td>
<td>2.83%</td>
<td>29.33%</td>
<td>1.89%</td>
<td>56.82%</td>
<td>11.28%</td>
</tr>
<tr>
<td></td>
<td>(6,225)</td>
<td>(18,241)</td>
<td>(661)</td>
<td>(13,241)</td>
<td>(38,368)</td>
</tr>
<tr>
<td><strong>in-</strong></td>
<td>31.91%</td>
<td>35.98%</td>
<td>44.59%</td>
<td>3.26%</td>
<td>31.99%</td>
</tr>
<tr>
<td></td>
<td>(70,079)</td>
<td>(22,372)</td>
<td>(15,580)</td>
<td>(760)</td>
<td>(108,791)</td>
</tr>
<tr>
<td><strong>non-</strong></td>
<td>7.42%</td>
<td>9.46%</td>
<td>0.84%</td>
<td>0.01%</td>
<td>6.61%</td>
</tr>
<tr>
<td></td>
<td>(16,287)</td>
<td>(5,884)</td>
<td>(292)</td>
<td>(3)</td>
<td>(22,466)</td>
</tr>
<tr>
<td><strong>un-</strong></td>
<td>57.01%</td>
<td>22.88%</td>
<td>52.48%</td>
<td>39.91%</td>
<td>49.13%</td>
</tr>
<tr>
<td></td>
<td>(125,210)</td>
<td>(14,231)</td>
<td>(18,340)</td>
<td>(9,301)</td>
<td>(167,082)</td>
</tr>
</tbody>
</table>

The data used for the current analysis is the same as that in Section 4.3.2.1 (Table 4.17) which examines the token distribution from the perspective of word classes. However, Table 4.19 presents information from the perspective of word class percentages over all negatively-prefixed tokens, and shows how likely tokens belonging to each word class are to include the negative prefixes. The column on the right demonstrates that if a token includes any of the prefixes, it is most likely to be **un-**, and least likely to be the **a-** prefix. This is also the case for adjectives and
adverbs, but for nouns, *in-* is the most common prefix, whereas for verbs, *dis-* is the most favoured. Other preferences can also be seen among the other word classes. As with types, verbs rarely incorporate *non-* and never take the *a-* prefix. To some extent, adjectives have a tendency to avoid *non-*, but particular shun *dis-* and *a-* respectively. Nouns have a tendency to avoid incorporating the *a-* prefix, and in the case of adverbs, *dis-*, *non-* and especially *a-* are the least likely prefixes.

4.3.2.3 Negatively-prefixed TOKENS compared with the whole BNC according to word class

The aim of this section is to present data which relates to the word class distributions of negatively-prefixed tokens and compare them with that of the BNC as a whole. This is to ascertain whether a specific negative prefix is more common in a particular word class, and, ultimately, to ascertain the degree to which word class affects the productivity of each of the negative prefixes. The total number of negatively-prefixed (adjectival, adverbial, nominal and verbal) tokens in the database amounts to 340,054. In order to discover the total number of adjectival, adverbial, nominal and verbal tokens in the whole of the BNC, Leech et al’s *Word Frequencies* (2001) was consulted. Leech et al classify the whole of the BNC corpus according to word class and provide approximate frequency figures per million tags (2001, List 6.1.1, p295). From this list, the approximate total frequencies were calculated for each of the four word classes¹⁴, which are shown in Table 4.20, along with the percentages per word class and the word class distribution of the negatively-prefixed tokens.

¹⁴ The BNC word class tags included were: JJ, JJR and JJT for adjectives; ND1, NN, NN1, NN2, NNA, NNB, NNL1, NNL2, NNO, NNO2, NNT1, NNT2, NNU, NNU1, NNU2, NP, NP1, NP2, NPD1, NPD2 and NPM1 for nouns; RA, REX, RG, RGQ, RGQV, RGR, RGT, RL, RP, RPK, RR, RRQ, RRQV, RRR, RRT and RT for adverbs; and VB0, VBDR, VBDZ, VBG, VBI, VBM, VBN, VBR, VBJ, VDO, VDD, VDG, VDL, VDN, VDZ, VH0, VHD, VH1, VH2, VHJ, VH1, VH2, VM, VMK, VV0, VVD, VVG, VVGK, VVI, VVN, VVNL and VVZ for verbs.
Table 4.20. Word class distribution percentages: negatively-prefixed TOKENS vs. whole BNC

<table>
<thead>
<tr>
<th>Word class</th>
<th>Total tokens for whole BNC (54,487,600)</th>
<th>Total tokens for all negatively-prefixed (340,054)</th>
<th>Total a-tokens (3,347)</th>
<th>Total dis-tokens (38,368)</th>
<th>Total in-tokens (108,791)</th>
<th>Total non-tokens (22,466)</th>
<th>Total un-tokens (167,082)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives</td>
<td>10.78% (5,876,100)</td>
<td>63.85% (219,618)</td>
<td>54.29% (1,817)</td>
<td>16.23% (6,225)</td>
<td>64.42% (70,079)</td>
<td>72.49% (16,287)</td>
<td>74.94% (125,210)</td>
</tr>
<tr>
<td>Nouns</td>
<td>39.50% (21,520,500)</td>
<td>18.08% (62,188)</td>
<td>43.62% (1,460)</td>
<td>47.54% (18,241)</td>
<td>20.56% (22,372)</td>
<td>26.20% (5,884)</td>
<td>08.52% (14,231)</td>
</tr>
<tr>
<td>Adverbs</td>
<td>12.72% (6,930,700)</td>
<td>10.16% (34,943)</td>
<td>2.09% (70)</td>
<td>1.72% (661)</td>
<td>14.32% (15,580)</td>
<td>1.30% (292)</td>
<td>10.98% (18,340)</td>
</tr>
<tr>
<td>Verbs</td>
<td>37.00% (20,160,390)</td>
<td>6.78% (23,305)</td>
<td>0% (0)</td>
<td>34.51% (13,241)</td>
<td>00.70% (760)</td>
<td>0.01% (3)</td>
<td>05.56% (9,301)</td>
</tr>
</tbody>
</table>

Table 4.20 (tokens) mirrors Table 4.16 (types) to some degree in that though nouns have the highest frequency across the whole BNC, it is adjectives that dominate among the negative prefixes. Nouns have similar percentages for both negatively-prefixed types and tokens (23.71% and 18.08% respectively), and likewise, adjectives have extremely similar percentages for negatively-prefixed types and tokens (65.64% and 63.85%). Adverbs are slightly more common in the BNC as a whole than their representation in negatively-prefixed tokens, although the percentages are very similar (12.72% and 10.16% respectively). Verbs are particularly well represented in the BNC in terms of token frequency (37.00%) compared to that of negatively-prefixed tokens as a whole (6.78%). The implications of the findings relating to the token frequencies are discussed in the following sections.

4.3.3 Summary of the role of word class in morphological productivity

The current section has presented data about word class in terms of type and token frequencies. It was important to engage with more refined data, since the individual

---

15 Total adjectival, nominal, adverbial and verbal tokens.
16 Excluding the prepositional tokens of unlike.
negative prefixes are not necessarily alike in the way they function and, therefore, a more fine-grained analysis can be developed; the current section has facilitated this. The aim of the section was to present the frequencies which relate to the word class distribution of the negative prefixes in order to discover whether each prefix has a predisposition towards combining with words from certain word classes, or indeed whether the individual prefixes have a tendency to shun certain classes.

Clearly, inclusion in a particular word class has an effect on the negative prefixes. Indeed, the effect is twofold, in that, while word class has an effect on the frequency distribution of the negative prefixes, likewise, the presence of negative prefixes affects the word class distribution (as can be seen from the word class ranking comparison with the BNC). The various frequencies presented in the current section are carried forward into Section 4.3.5 and facilitate productivity calculations according to word class.

4.3.4 Two case studies: word class TYPES and TOKENS for specific prefixes

Up to this point in this chapter, a number of methodologies have been applied to the negative prefixes, in addition to an investigation of the relevance of hapax legomena. The previous sections have examined the role of word class in terms of its potential effect on morphological productivity. The current section examines two of the methods applied to date - the hapax/type and the hapax/token method - with regard to specific examples of prefixes in particular word classes. The purpose of this is to highlight some of the issues of the methods applied, the notion of hapax legomena, the influence of word class and the idiosyncrasies of the individual negative prefixes. It is also to show how some of the statistical information presented above can be implemented in practice, particularly in light of the new evidence presented about the degree to which hapaxes can be considered neologisms.
4.3.4.1 *A* prefixed adjectival types and the hapax/type ratio

Table 4.16 showed the distribution of the negatively-prefixed types according to word class. Negatively-prefixed *a*- adjectives constitute the smallest percentage in terms of adjectival types, and it may, therefore, be interesting to analyse *a*- adjectival types according to Baayen and Lieber’s (1991) hapax/type ratio. In terms of negatively-prefixed adjectives in the database, *a*- is made up of 88 different types; 31 of these are hapaxes. Thus, according to Baayen’s methods, the hapax/type ratio for *a*- prefixed adjectival types is 0.35. The equation of hapaxes and neologisms, however, has now been brought into question (Section 4.2), and indeed, the adjectival *a*- hapaxes have been examined by means of the OED to clarify their status. Table 4.21 shows that of the 31 adjectival *a*- hapaxes, two were coined in the 1600s, fifteen are from the 1800s, and six from the 1900s, which makes a total of 23 adjectival *a*- hapaxes which would certainly not be classified as neologisms, since they are long established words of the language.

<table>
<thead>
<tr>
<th><em>a</em>- Adjectival Hapax</th>
<th>Negative Form Coinage Date</th>
<th>Neologism?</th>
<th><em>a</em>- Adjectival Hapax</th>
<th>Negative Form Coinage Date</th>
<th>Neologism?</th>
</tr>
</thead>
<tbody>
<tr>
<td>areligious</td>
<td></td>
<td>Yes</td>
<td>anosmic</td>
<td>1875</td>
<td>No</td>
</tr>
<tr>
<td>apersonal</td>
<td></td>
<td>Yes</td>
<td>asyntactic</td>
<td>1880</td>
<td>No</td>
</tr>
<tr>
<td>arhythmic</td>
<td></td>
<td>Yes</td>
<td>amitotic</td>
<td>1888</td>
<td>No</td>
</tr>
<tr>
<td>atypic</td>
<td>1601</td>
<td>No</td>
<td>anastigmatic</td>
<td>1890</td>
<td>No</td>
</tr>
<tr>
<td>ahelical</td>
<td></td>
<td>Yes</td>
<td>aniconic</td>
<td>1892</td>
<td>No</td>
</tr>
<tr>
<td>azygos</td>
<td>1646</td>
<td>No</td>
<td>anhedral</td>
<td>1896</td>
<td>No</td>
</tr>
<tr>
<td>aglandular</td>
<td></td>
<td>Yes</td>
<td>anhedralled</td>
<td>1896</td>
<td>No</td>
</tr>
<tr>
<td>abaconian</td>
<td></td>
<td>Yes</td>
<td>azonal</td>
<td>1896</td>
<td>No</td>
</tr>
<tr>
<td>anisotenic</td>
<td></td>
<td>Yes</td>
<td>achondritic</td>
<td>1917</td>
<td>No</td>
</tr>
<tr>
<td>anisopteran</td>
<td>18??</td>
<td>No</td>
<td>aspheric</td>
<td>1923</td>
<td>No</td>
</tr>
<tr>
<td>anormal</td>
<td>1835</td>
<td>No</td>
<td>atraumatic</td>
<td>1934</td>
<td>No</td>
</tr>
<tr>
<td>anuric</td>
<td>1838</td>
<td>No</td>
<td>anicteric</td>
<td>1935</td>
<td>No</td>
</tr>
<tr>
<td>akinetic</td>
<td>1841</td>
<td>No</td>
<td>aceramic</td>
<td>1961</td>
<td>No</td>
</tr>
<tr>
<td>astylar</td>
<td>1842</td>
<td>No</td>
<td>anoxygencic</td>
<td>1975</td>
<td>No</td>
</tr>
<tr>
<td><em>a</em>-centralized</td>
<td>1842</td>
<td>No</td>
<td>abio-energetic</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
Potentially, therefore, eight of the hapaxes (*areligious, apersonal, arhythmical, ahelical, aglandular, abaconian, anisotenic* and *abio-energetic*) may be classed as neologisms, as they do not appear in their negated forms in the *OED*. The final example, *abio-energetic*, does not occur in the *OED* in either its positive or negative form\(^\text{17}\), although all the components which make up the word are present in the *OED* (*a-, bio- and energetic*). The significance of this is that the hapax/type productivity score (31/88=0.35) can be replaced with a more accurate neologism/type quotient (8/88=0.09).

### 4.3.4.2 *Dis*-prefixed verbal tokens and the hapax/token ratio

Table 4.18 showed the word class percentages for the negatively-prefixed tokens, and demonstrated that *dis*- prefixed tokens had the highest proportion of verbs of all the negatively-prefixed tokens, with over a third of all *dis*- tokens being verbs. Indeed, *dis*- demonstrated the highest proportion of negatively-prefixed verbal tokens than any of the other prefixes. On this basis, it may be interesting to analyse the *dis*- verbal tokens according to Baayen and Lieber’s (1991) hapax/token ratio.

*Dis*- verbs in the database consist of 13,241 tokens; 21 of these are hapaxes. The proportion of hapaxes per token, therefore, is measured as 0.002, which, compared to the calculations in Table 4.5 in Section 4.1.4, shows a relatively low hapax/token ratio for the prefix in terms of verbal tokens. Table 4.22 shows the *dis*- verbal hapaxes examined in the *OED* in order to assess their status as neologisms or not.

---

\(^{17}\)The word is recorded in the BNC context as being from a 1991 publication.
Table 4.22. Verbal Dis-Prefixed Hapaxes – Potential Neologisms

<table>
<thead>
<tr>
<th>dis- Verbal Hapax</th>
<th>Negative Form Coinage Date</th>
<th>Neologism?</th>
</tr>
</thead>
<tbody>
<tr>
<td>dis-represent</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>disnature</td>
<td>1450</td>
<td>No</td>
</tr>
<tr>
<td>disforest</td>
<td>1502</td>
<td>No</td>
</tr>
<tr>
<td>dispark</td>
<td>1538</td>
<td>No</td>
</tr>
<tr>
<td>disbeautify</td>
<td>1577</td>
<td>No</td>
</tr>
<tr>
<td>discountenance</td>
<td>1580</td>
<td>No</td>
</tr>
<tr>
<td>disinvite</td>
<td>1580</td>
<td>No</td>
</tr>
<tr>
<td>discommunicate</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>disencumber</td>
<td>1596</td>
<td>No</td>
</tr>
<tr>
<td>disentail</td>
<td>1641</td>
<td>No</td>
</tr>
<tr>
<td>disimprove</td>
<td>1642</td>
<td>No</td>
</tr>
<tr>
<td>disincine</td>
<td>1647</td>
<td>No</td>
</tr>
<tr>
<td>discorporate</td>
<td>1683</td>
<td>No</td>
</tr>
<tr>
<td>disincorporate</td>
<td>1697</td>
<td>No</td>
</tr>
<tr>
<td>disrate</td>
<td>1811</td>
<td>No</td>
</tr>
<tr>
<td>disentwine</td>
<td>1814</td>
<td>No</td>
</tr>
<tr>
<td>disendow</td>
<td>1861</td>
<td>No</td>
</tr>
<tr>
<td>disunify</td>
<td>1891</td>
<td>No</td>
</tr>
<tr>
<td>disinfest</td>
<td>1920</td>
<td>No</td>
</tr>
<tr>
<td>disinhabit</td>
<td>1927</td>
<td>No</td>
</tr>
<tr>
<td>disproportionate</td>
<td>1934</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4.22 shows that of the 21 verbal dis- hapaxes, twelve appeared in the language between 1450 and 1700, and seven between 1800 and 1950. The final two examples, discommunicate and dis-represent, do not appear in their exact forms in the OED, though discommunicate does show similarities with discommune and discommunion, which are both dated from 1590. Dis-represent is not listed in the OED, though its positive verbal form, represent is shown as dating from 1380. These final two examples are perhaps the only examples which could possibly be considered as neologisms. This results in a neologism/token score of 0.0002 (2/13,241). This suggests that though dis- verbs are common in the language, they are not prone to the production of new examples. Indeed, Baayen & Leiber’s (1991) hapax/token measure would point towards dis- verbs being more productive than they actually are. I now turn to the application of the measures examined earlier (Measures A, B1 and C1) to discuss the negative prefixes in terms of word class productivity, and begin with the mean type frequency measure.
4.3.5 Measure A - Word class productivity – mean type frequency for each prefix

Baayen and Lieber’s (1991) mean type frequency measure, which was replicated in Section 4.1.3, and further discussed in Section 4.1.6, with regard to all the type and tokens in the database, can also be reproduced to ascertain the productivity of the various prefixes in terms of their inclusion in lexical items within the individual word classes. Recall that a low mean type frequency is said to associate with a high level of productivity. The results can be seen in Table 4.23 below.

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>(1,817 / 88)</td>
<td>(70 / 16)</td>
<td>(1,460 / 46)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td>20.648</td>
<td>4.375</td>
<td>31.739</td>
<td>N/A</td>
</tr>
<tr>
<td>dis-</td>
<td>(6,225 / 143)</td>
<td>(661 / 27)</td>
<td>(18,241 / 166)</td>
<td>(13,241 / 93)</td>
</tr>
<tr>
<td></td>
<td>43.531</td>
<td>24.481</td>
<td>109.885</td>
<td>142.376</td>
</tr>
<tr>
<td>in-</td>
<td>(70,079 / 351)</td>
<td>(15,580 / 193)</td>
<td>(22,372 / 358)</td>
<td>(760 / 20)</td>
</tr>
<tr>
<td></td>
<td>199.655</td>
<td>80.725</td>
<td>62.491</td>
<td>38.000</td>
</tr>
<tr>
<td>non-</td>
<td>(16,287 / 2,525)</td>
<td>(292 / 70)</td>
<td>(5,884 / 1,188)</td>
<td>(3 / 1)</td>
</tr>
<tr>
<td></td>
<td>6.450</td>
<td>4.171</td>
<td>4.952</td>
<td>N/A</td>
</tr>
<tr>
<td>un-</td>
<td>(125,210 / 2,879)</td>
<td>(18,340 / 338)</td>
<td>(14,231 / 404)</td>
<td>(9,301 / 214)</td>
</tr>
<tr>
<td></td>
<td>43.490</td>
<td>54.260</td>
<td>35.225</td>
<td>43.462</td>
</tr>
<tr>
<td>All prefixes</td>
<td>(219,618 / 5,986)</td>
<td>(34,943 / 644)</td>
<td>(62,188 / 2,162)</td>
<td>(23,305 / 328)</td>
</tr>
<tr>
<td></td>
<td>36.689</td>
<td>54.259</td>
<td>28.764</td>
<td>71.052</td>
</tr>
</tbody>
</table>

By calculating the totals for each word class, it can be demonstrated whether negative prefixing is a more productive process when adjectives, nouns, adverbs or verbs are created. Table 5.23, therefore, shows that negative prefixing is the most productive when nouns are created, whereas they are at their least productive when adjectives are formed.

All of the negating prefixes have different levels of productivity across the different word classes, and these productivity levels vary markedly between the

---

18 Where a prefix has no type presence in a particular word class or where token frequency is less than 10, these are discounted for the purposes of the calculations.
prefixes. Taking the prefix dis-, for example, the lowest mean type frequency for dis-is 24.481 in the case of adverbs. However, in terms of verbs, dis- has one of the highest mean type frequencies of all the prefixes over all the word classes with 142.376, which by this measure indicates that dis- is not productive in the case of verbs. This demonstrates that productivity is sensitive to, and affected by, word class, a point which has never been made before.

Although Baayen and Renouf (1996) begin to superficially examine word class, and claim for example that “adjectivizing -ly is much less productive than adverbializing -ly” (82), and that the prefix un- is more commonly attached to adjectives than to verbs (85), it is not an issue that they subsequently pursue and, indeed, it only receives a cursory mention in their conclusion. Baayen and Renouf’s (1996) belief is that “(c)oncentrating on the hapax legomena as prime instantiations of lexical innovations” demonstrates the productivity of a process (82). However, their methodology, in this instance, involves a simple count of the number of hapaxes in their corpus which are suffixed by either adjectival -ly or adverbial -ly. Similarly, they count the number of adjectival hapaxes prefixed with un- compared to the number of verbal hapaxes prefixed with un-. Involving straightforward counting methods, Baayen and Renouf’s (1996) methodology appears not to make use of Baayen’s earlier measures (1989) or those proposed by Baayen and Lieber in 1991, and is somewhat different than the measures examined by the current research. Clearly, simple hapax-counting does not measure productivity and, as this current research has shown, hapaxes do not constitute lexical innovation. It is interesting to note that the issue of word class is not raised again by Baayen and his colleagues, and the current study is, in fact, the first to examine the notion in any detail.
Returning to Measure A, the mean type frequency measure, across all the word classes, negatively-prefixed nouns are demonstrated as the most productive class, whereas verbs are the least productive word class for the negative prefixes. The overall productivity rankings in terms of the mean type frequency measure are shown in Table 4.24 below. What is perhaps surprising is that it is adverbs prefixed with *non*- which are the most productive of all the word classes for all the prefixes, according to this measure. Section 3.5.4 showed that *non*- constructions have a tendency towards binary or non-gradable opposition; in other words, contradictory negation, and this would tend to deny the need for any considerable number of adverbs, which would typically serve to modify a description, an action or state, or a whole statement.

Table 4.24. Measure A - Word class productivity rankings based on a token/type measure for each prefix

| Prefix / word class | 1 non- adverbs | 2 a- adverbs | 3 non- nouns | 4 non- adjectives | 5 a- adjectives | 6 dis- adverbs | 7 a- nouns | 8 un- nouns | 9 in- verbs | 10 un- verbs | 11 un- adjectives | 12 dis- adjectives | 13 un- adverbs | 14 in- nouns | 15 in- adverbs | 16 dis- nouns | 17 dis- verbs | 18 in- adjectives |
According to Jones (2002), these are ‘Coordinated Antonyms’ of the “X and Y” framework (63), which “signal inclusiveness or exhaustiveness of scale” (74). Examples of these co-occurring non-adverbs are shown in the BNC contexts below.

(i) There are, however, great differences in linguistic style between **academically** and **non-academically** oriented children.

(ii) At this stage, members of the group are still tolerant but are signalling, **verbally** and **non-verbally**, that the deviation must cease.

(iii) The net change in height centile compared with diagnosis was comparable for **surgically** and **non-surgically** treated patients....

(iv) The MRC provides grants for **medically** and **non-medically** qualified graduates....

Apart from dis-adverbs, both dis- and in- word classes appear in the lower section of the rankings, which indicates their general lower level of productivity across all word classes. By this measure, in- adjectives are demonstrated as the least productive prefix of all the word classes.

**4.3.6 Applying the neologism/token (Measure B1) and neologism/type (Measure C1) ratios to the word class distributions**

The next two measures to be applied (the B and C Measures) serve to compare with and support the previous type frequency measure (Measure A). To calculate the hapax/token and hapax/type measures for each word class, the number of hapaxes from the database for each of the negative prefixes for each of the four word classes was taken. The distribution of hapaxes across the various word classes is shown in Table 4.25, which enables the hapax/token (Measure B) and hapax/type (Measure C) ratios to be determined. However, it is now possible to use the more refined measures (Measure B1 – neologism/token method and C1 – neologism/type method)
developed in Section 4.2.4 and apply these to the word class distributions for the negative prefixes.

Table 4.25. Distribution of the hapaxes for each word class for each prefix

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Verbs</th>
<th>Total Number of Hapaxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>31</td>
<td>6</td>
<td>13</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>dis-</td>
<td>37</td>
<td>8</td>
<td>46</td>
<td>21</td>
<td>112</td>
</tr>
<tr>
<td>in-</td>
<td>58</td>
<td>32</td>
<td>91</td>
<td>8</td>
<td>189</td>
</tr>
<tr>
<td>non-</td>
<td>1,265</td>
<td>40</td>
<td>693</td>
<td>0</td>
<td>1,998</td>
</tr>
<tr>
<td>un-</td>
<td>1,088</td>
<td>112</td>
<td>199</td>
<td>69</td>
<td>1,468</td>
</tr>
</tbody>
</table>

In the following two tables (Table 4.26 and 4.27), the bracketed figures refer to the number of hapaxes for that prefix for that word class, which is then multiplied by the decimalised percentage previously calculated for that particular prefix (a- 0.20; dis- 0.12; in- 0.06; non- 0.85; un- 0.23), which is then divided by the number of tokens (B1) or types (C1) for that prefix for that word class. Thus, the subsequent calculations, therefore, demonstrate the productivity of the prefixes in each word class according to the neologism/token (B1) and the neologism/type (C1) measures. First, the neologism/token calculations are shown in Tables 4.26 below.

Table 4.26. Measure B1 - Word classes – neologism/token ratio

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Verbs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>0.003412</td>
<td>0.017143</td>
<td>0.001781</td>
<td>(0 / 0)</td>
<td>N/A</td>
</tr>
<tr>
<td>dis-</td>
<td>0.000713</td>
<td>0.001452</td>
<td>0.000303</td>
<td>(21 x 0.12 / 13,241)</td>
<td>0.000190</td>
</tr>
<tr>
<td>in-</td>
<td>0.000050</td>
<td>0.00123</td>
<td>0.000244</td>
<td>(8 x 0.06 / 760)</td>
<td>0.000632</td>
</tr>
<tr>
<td>non-</td>
<td>0.066018</td>
<td>0.116438</td>
<td>0.100110</td>
<td>(0 x 0.85 / 3)</td>
<td>N/A</td>
</tr>
<tr>
<td>un-</td>
<td>0.001998</td>
<td>0.001404</td>
<td>0.003216</td>
<td>(69 x 0.23 / 9,301)</td>
<td>0.001706</td>
</tr>
</tbody>
</table>
Taking the negative-prefix/word-class combinations, non- adverbs are the most productive, with non- nouns and non- adjectives in second and third places respectively. In- prefixed adjectives are the least productive of all the prefixes across all the word classes according to the neologism/token measure.

The neologism/type calculations are shown in Tables 4.27 below. Interestingly, although un- nouns are not the most productive in terms of nouns as a whole, they do constitute the highest proportion of all the un- types, which again supports Horn’s (2002) view of the current increase in the productivity of novel un- nouns.

Table 4.27. Measure C1 - Word classes - neologism/type ratio

<table>
<thead>
<tr>
<th></th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Nouns</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a-</td>
<td>0.070455</td>
<td>0.075000</td>
<td>0.056522</td>
<td>N/A</td>
</tr>
<tr>
<td>dis-</td>
<td>0.031049</td>
<td>0.035556</td>
<td>0.033253</td>
<td>0.027097</td>
</tr>
<tr>
<td>in-</td>
<td>0.009915</td>
<td>0.009948</td>
<td>0.015251</td>
<td>0.024000</td>
</tr>
<tr>
<td>non-</td>
<td>0.425842</td>
<td>0.485714</td>
<td>0.495833</td>
<td>N/A</td>
</tr>
<tr>
<td>un-</td>
<td>0.086919</td>
<td>0.076213</td>
<td>0.113292</td>
<td>0.074159</td>
</tr>
</tbody>
</table>

Taking the negative-prefix/word-class combinations, non- nouns are the most productive by this measure, closely followed by non- adverbs and non- adjectives, whereas in- prefixed adjectives are the least productive of all the prefixes across all the word classes. In the case of negatively-prefixed verbs, these are the least likely word class to be encountered in terms of new types, and even when new verbal types do occur, they are not likely to be used often in the language.

Following the view of Van Marle (1992), the current research has also adopted an approach which uses types as a variable. However, this research has
enabled more sophisticated measures to be developed. Rather than the hapax/type measure proposed by Baayen and Lieber (1991), this study has developed a procedure to identify the genuine neologisms in a list of hapaxes and can, therefore, propose the neologism/type measure as the most appropriate method for measuring morphological productivity. Having examined the role of word class on productivity, the following section investigates another potential influence on individual negative prefixes, that of domain.

4.4 Which negative prefix favours which domain?

The previous section investigated the effect of word class on the negative prefixes. A further aim of this research is to establish whether the negative prefixes have a preference for particular genres. In other words, this section of the study determines whether the prefixes are more prevalent in certain domains. This section is therefore intended to provide a snapshot of the different domains in which individual negating prefixes are most likely to occur, with a view to prompting even more fine-grained calculations of productivity in future studies.

In order to address the question, the same sample of 751 hapaxes used in Section 4.2.2 (approximately 20% of the 3,817 hapaxes in the database) was again examined, and each of the 751 entries was analysed for context and bibliographic information that relates to domain. This was carried out by means of the BNCweb (CQP-Edition), and the text domain was recorded for each of the 751 hapaxes. These figures were also subsequently compared to the domain distribution for the whole of the BNC (http://www.natecorp.ox.ac.uk/docs/userManual/bncIndex.html and Leech et al, 2001:2). The distribution of the different hapaxes for each negative prefix across the different domains is shown in Table 4.28 below.
Table 4.28. Domain distribution of hapax sample for each negative prefix and for the BNC as a whole

<table>
<thead>
<tr>
<th>Domains</th>
<th>a-</th>
<th>dis-</th>
<th>in-</th>
<th>non-</th>
<th>un-</th>
<th>all prefixes</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Science</td>
<td>22.00%</td>
<td>17.86%</td>
<td>12.70%</td>
<td>22.00%</td>
<td>11.00%</td>
<td>16.11%</td>
<td>13.32%</td>
</tr>
<tr>
<td>World Affairs</td>
<td>8.00%</td>
<td>21.43%</td>
<td>14.29%</td>
<td>12.50%</td>
<td>13.00%</td>
<td>14.12%</td>
<td>16.55%</td>
</tr>
<tr>
<td>Commerce</td>
<td>0.00%</td>
<td>9.82%</td>
<td>3.17%</td>
<td>7.50%</td>
<td>8.00%</td>
<td>6.39%</td>
<td>7.14%</td>
</tr>
<tr>
<td>Leisure</td>
<td>4.00%</td>
<td>5.36%</td>
<td>7.94%</td>
<td>9.00%</td>
<td>8.50%</td>
<td>7.72%</td>
<td>10.03%</td>
</tr>
<tr>
<td>Natural Science</td>
<td>26.00%</td>
<td>2.68%</td>
<td>5.82%</td>
<td>12.50%</td>
<td>7.50%</td>
<td>8.92%</td>
<td>3.76%</td>
</tr>
<tr>
<td>Imaginative</td>
<td>6.00%</td>
<td>11.61%</td>
<td>13.23%</td>
<td>4.50%</td>
<td>27.00%</td>
<td>13.85%</td>
<td>19.72%</td>
</tr>
<tr>
<td>Belief &amp; Thought</td>
<td>4.00%</td>
<td>4.46%</td>
<td>8.47%</td>
<td>4.50%</td>
<td>4.00%</td>
<td>5.33%</td>
<td>3.06%</td>
</tr>
<tr>
<td>Arts</td>
<td>6.00%</td>
<td>16.96%</td>
<td>15.87%</td>
<td>12.50%</td>
<td>13.50%</td>
<td>13.85%</td>
<td>7.27%</td>
</tr>
<tr>
<td>Applied Science</td>
<td>22.00%</td>
<td>5.36%</td>
<td>8.99%</td>
<td>13.50%</td>
<td>6.50%</td>
<td>9.85%</td>
<td>7.40%</td>
</tr>
<tr>
<td>Unclassified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.74%</td>
</tr>
<tr>
<td>All written</td>
<td>98.00%</td>
<td>95.54%</td>
<td>90.48%</td>
<td>98.50%</td>
<td>99.00%</td>
<td>96.14%</td>
<td>89.99%</td>
</tr>
</tbody>
</table>

| Spoken           |       |        |        |        |        |              |         |
| Education        | 2.00% | 2.68%  | 2.64%  | 1.00%  |        | 1.67%        | 2.06%   |
| Business         | 0.89% | 1.06%  | 1.06%  | 0.50%  |        | 0.49%        | 2.15%   |
| Leisure          | 0.89% | 1.06%  | 1.06%  | 0.50%  |        | 0.49%        | 2.37%   |
| Public           | 4.23% |        |        |        |        | 0.85%        | 2.19%   |
| Conversation     | 0.53% | 0.05%  |        |        |        | 0.12%        | 1.24%   |
| All spoken       | 2.00% | 4.46%  | 9.52%  | 1.50%  | 1.00%  | 3.86%        | 10.01%  |

(BNC percentages taken from The BNC Handbook, 1998:29 and 32\(^{19}\)).

Overall, “(t)he BNC contains approximately 90 per cent written data and 10 per cent spoken data” (Leech et al, 2001:1), and this is broken down into the various domains in Table 4.28. Negatively-prefixed words are more likely to occur in written

\(^{19}\)Percentage figures have been adapted based on the written component of the BNC which is 90% of the whole corpus, and the spoken component which is 10% of the whole corpus.
language than in the general distribution of the *BNC* as a whole (approximately 90%). As Tottie pointed out before the widespread use of corpora, negatively-prefixed words (affixal negation) are more likely to occur in written language, whereas sententially (non-affixal negation) is more likely to be found in spoken language (1980:104). The table shows, written negatively-prefixed words are more likely to be in the areas of Social Science, World Affairs, Imaginative writing or the Arts, but as a whole, the negative prefixes tend to avoid Belief and Thought.

With individual prefixes, some marked trends appear. For example, *a-* prefixed words are more likely to be in the field of Natural Science, and *non-* favours Social Science. *Dis-* and *in-* appear to favour World Affairs and the Arts respectively, whereas *un-* seems to be more common in Imaginative writing but shuns the language of Belief and Thought. *Non-* also particularly tends to avoid Belief and Thought and Imaginative writing.

In terms of spoken domains, Educational settings are the most popular for the negative prefixes, though this is a lower proportion than for the *BNC* as a whole, and *un-* appears to avoid Educational settings all together. *In-* appears in all spoken domains, though is most popular in Public contexts. *Dis-, in-* and *un-* are the only prefixes which appear in Business and Leisure contexts, and *in-* and *non-* are the only prefixes to have some presence in Conversational settings.

However, the most striking aspect of Table 4.28 is the degree to which the negating prefixes shun spoken language as a whole in favour of written language. All five of the negative prefixes do this to some extent, though the preference towards writing over speech shown by *in-* is negligible. What is most interesting is that those prefixes shown to be the most productive in the formulae devised by Baayen and his co-researchers are those that are least common in spoken English. Indeed, the
ranking for spoken English is a perfect reversal of the one typically found in the productivity measures presented earlier in this chapter, whereby *in-* is the most common in spoken language, followed by *dis-*, *a-*, *un-* and *non-*. The most likely explanation for this is that productivity is more a feature of written language than spoken language, and that those prefixes that are not currently used in new coinages – in other words, those that occur primarily as part of familiar, often high frequency words – tend to be used more in speech.

It is difficult to draw more concrete conclusions based on such small amounts of data which relate to domain preference, but initial observations indicate some interesting patterns beginning to emerge in the distribution of the written/spoken data. Whilst the research into the negative prefixes in terms of their domain preferences and their written/spoken distribution has only been an indicative measure, it is apparent that domain also has an effect on the choice of one prefix over another. Certainly, this is an area which has been overlooked previously and may well benefit from further investigation in the future.

### 4.5 Chapter Summary

This chapter has used *BNC* data to investigate negative prefixes in order to reappraise the previously-tried methodologies for measuring morphological productivity. The chapter initially examined methods which involve types and also those which involve tokens, but found that straightforward type-counting or token-counting measures did not constitute an adequate measure of productivity, nor did they give any indication of the current status of a morphological process. The type/token measure was next examined, to ascertain the mean token frequency for each of the negative prefixes. Baayen and Lieber propose that low mean token
frequency is associated with higher productivity (1991:809), and the method clearly brings in a more two-dimensional perspective to the equation. The chapter went on to apply hapax/token and hapax/type methods to the negative prefixes, and it is proposed that these methods are perhaps a move closer towards a more realistic measure of productivity.

It was also determined, however, that hapaxes require more detailed examination than has previously been attempted, in order to re-assess their value as an indicator of morphological productivity. As a result of this closer investigation, hapaxes were categorised, since it was shown that they do not necessarily equate with neologisms. Word class data was also generated in order to assess the influence of word class on the distribution of prefixes, and finally, the negative prefixes were assessed in terms of domain analysis, whereby the distribution of the negative prefixes was examined across all written and all spoken domains.

The chapter has used the negative prefixes as a case study to examine a number of variables, and in the case of one of the variables - hapax legomena - the phenomenon has been investigated in greater detail. The purpose of the chapter has been to gather the appropriate frequency data, and to set out the information that is required in order to move towards more sophisticated measures, and to interpret and discuss the statistical evidence. A number of issues have been addressed and a number of findings have been reported:

- five different methods of measuring productivity have been investigated (type; token; type/token ratios; hapax/token; and hapax/type ratios) and it has been confirmed that straightforward type- or token-counting methods reveal little about productivity;
- productivity measures which involve more than one variable potentially offer more realistic outcomes, since they provide a two-dimensional perspective;
• hapax legomena are a useful phenomenon in measuring productivity, and provide access to neologisms, but are not, in themselves, a reflection of productivity, and require more sophisticated analysis;

• hapax legomena typically include a percentage of neologisms, but this percentage varies significantly from prefix to prefix;

• productivity measures which introduce actual newly-coined words into the equation provide more sophisticated means of assessing productivity;

• across all the measures applied, non- emerges as the most productive of the negative prefixes, with dis- and in- the least productive;

• productivity is sensitive to, and affected by, word class, and it has been demonstrated that negative prefixing is a more productive process when nouns are created, and least productive when verbs are formed, though patterns vary from prefix to prefix.

The chapter has discussed various theories and approaches to measuring morphological productivity. Particularly, it has focussed on the measures proposed by Baayen and Lieber (1991) in which they promote the idea (originally put forward by Baayen, 1989) of hapax legomena in measuring productivity. The chapter has then gone on to refine the hapax/token and hapax/type measures proposed by Baayen, by developing neologism/token and neologism/type measures.

The prefix non- has, via all the methods applied, consistently been demonstrated as the most productive of all the negative prefixes. It is proposed by the current research that the high level of productivity of non- is partially due to its neutral status as a negative prefix and also as a result of its lack of pejorative connotations. Even in the pre-corpus era, Zimmer (1964) proposed that contradictory negation is a more productive process than contrary negation (32), and indeed, non- is the only one of the five negative prefixes which creates purely contradictory...
negation. Baayen and Lieber also claim that an affix is productive if a considerable number of its types are hapaxes, and over 50% of non-types in the database are hapaxes.

This research also submits that the apparent productivity of the various negative prefixes may relate to the date of their introduction into the language; how the prefixes were utilised at that time; and how the prefixes might be used today, based on restrictions, the introduction of new senses, etc. In other words, productivity may be affected by a number of diachronic factors.

In terms of the various productivity measures applied in this chapter, this research, to some extent, adopts the view of Van Marle (1992), and takes up the argument in favour of the use of type frequency in measuring productivity. That is not to say that the notion of token frequency is dismissed by this research, despite the fact that Van Marle questions the validity of token frequency in a measure of productivity, and argues that “(o)nce a word is coined, the frequency in use of that word, [...] is more or less irrelevant to the degree of productivity of that rule.” (1992:156). More, this study favours the use of the ‘type’ variable, but views other measures as supporting evidence. Van Marle states that Baayen’s work concerns itself with ‘likelihood’, rather than the actual ‘ability’ of a process to be productive (1992:152) and, on this basis, he challenges Baayen’s view with regards the notion of hapaxes, and proposes that they need further and more sophisticated investigation. This study has aimed to be that further and more sophisticated investigation. Both this research and Van Marle (1992) do not wish to diminish the outstanding and far-reaching contribution which Baayen has made to the notion of morphological productivity, but as Van Marle claims:
....the impressive sophistication of the statistical and mathematical procedures that are used [by Baayen] is not always paralleled by a similar sophistication of the linguistic methods and findings.

(1992:160)

Despite his opinion of hapaxes, Van Marle maintains that the relationship “between hapaxes and types is of more interest to productivity” than that between hapaxes and tokens (1992:158), and this research supports that view. In examining the hapaxes in greater detail, this research has sought to identify the proportion of neologisms contained within a list of hapaxes, and to utilise these neologisms in a more fine-grained measure of productivity. While it has been demonstrated that hapaxes do not equate to neologisms, and that indeed, most hapaxes are not neologisms, it has also been shown that the percentage of neologisms in a list of hapaxes vary considerably from affix to affix. A list of hapaxes needs to be filtered in order to leave a ‘useful’ list which is, in fact, a list of neologisms.

With regard to the word class inclusion of the negative prefixes, the chapter has demonstrated that negative prefixes are affected by word class and, in turn, they too affect the word classes to which they belong. Again, this is a useful finding in that it demonstrates that the negative prefixes function very differently to one another and, likewise, it also shows that productivity measures need to be tailored to suit the different prefixes.

This thesis has re-assessed the methodologies proposed by a number of corpus linguists to measure morphological productivity. Its purpose has been to replicate, complement and refine earlier, but often highly original, studies by means of the use of negative prefixation as a case study. The following chapter re-examines the aims of this research, and discusses how these aims have been met and how the
notion of morphological productivity has been moved forward, and has also considered how subsequent research may be developed.
Chapter 5

Conclusion

The previous chapters of this thesis have examined productivity from the perspective of a case study based on five negative prefixes of English. This concluding chapter evaluates the results of this study by synthesising and interpreting the findings to determine how our understanding of productivity has moved on. To this end, the chapter re-examines the aims of the research, reviews the findings of the research, and discusses its limitations, in addition to presenting suggestions for possible future research.

The capability of language to produce new complex words is a vital ingredient in keeping a language alive. Words of a language may become rare, or may even become obsolete; other words may develop new meanings, or meanings may be lost. But equally and critically important to any language is its ability to create new words. As Katamba suggests, the words of a language are not simply those we store in the mental lexicon or list in dictionaries:

....although by and large the words of a language are listable in dictionaries, it is not possible to list all of them. Speakers actively manufacture words. The lexicon is not just a vast lexical warehouse, it is also a production line where words are made in limitless quantities.  

(1994:153)

The phenomenon of productivity is, therefore, an indispensable feature of any language in order to ensure its development, and indeed its very existence. Thus, the components which make up complex words are the vital ingredients which allow a
language to be dynamic. In the past, judgements regarding the productivity of these components were, by necessity, based on intuition but, due to the introduction of large-scale electronic corpora, research into language productivity has been able to take a more empirical approach, based on genuine examples of language; that is, language which has actually occurred.

Although there has been considerable research into the notion of productivity, and despite the fact that we all engage in language productivity to some extent, it remains a complex issue. As Bauer argues, “productivity remains one of the most contested areas in the study of word-formation” (1983:62). Sometimes we produce new words knowingly for effect; sometimes unwittingly; and sometimes to create a word to name or describe a new concept or entity. The creation of a new word typically, though not exclusively, involves the use of existing components with which we are familiar, in a combination which has, to that point, never been previously made. Especially in a time of rapidly developing technology, there is an even greater need to create new words. Via the Internet and the media in general, we are constantly surrounded by new word formations which, although we may not have encountered before, we are, nevertheless, generally able to understand, due to our knowledge of these components and our knowledge of the world. The following, for instance, are very recent examples of negatively-prefixed neologisms and, thus, examples of productivity:

“Sadly, I'm still among the un-wiied.”
(www.metatalk.metafilter.com, 09/03/08)

“I've retold that joke like every time ungoogleability comes up in conversation, which is a lot because I'm in a band called Google Maps.”
(http://story-games.com, 02/07/09)
“I wish I could use it to send messages to my *non-iPhoning* friends!”

(www.iphoneappsplus.com, 19/11/09)

“Finding a *non-Googleable* source of quotes/jokes can make you a cool thinker.”

(http://minekey.com, 09/01/10)

“Britney Spears asks for *non-airbrushed* images to be released”.

(www.feministphilosophers.wordpress.com, 14/04/10)

The words are made up of base words which relate to contemporary technology, along with readily recognised suffixes and negative prefixes and, as a consequence, the meaning of all the words can be understood, despite the probability that the reader has never before come across these words.

There still appears to be no common agreement among linguists as to what exactly it is in language that is productive. For some linguists, it is the affixes themselves which are either productive or not; some believe it is the processes which may or may not be productive; others consider that it is the rules of the language which are productive or not; and others still that it is the combination of the various components involved which dictate the productivity. Bauer (2001) examines this variety of opinions and states that “(b)y and large these various descriptions can be seen as differing statements of a single phenomenon, with a greater or lesser degree of specificity.” (12). In other words, what is productive is not the main issue here but, as Di Sciullo and Williams propose, it relates more to the way in which real speakers use the language (1987:2). Thus, the process is not the main issue; but the output of that process. This research has, therefore, taken a combined view of productivity, whereby it is based on a rule-governed system which involves derivational processes,
uses base words and affixes to create new words, and utilises genuine evidence of language which real speakers have used.

Language productivity is the instrument by which a language changes and advances, and this research has sought to clarify the complex aspects pertaining to language productivity in order to take the notion a step forward. The rationale has not been to detract from the validity of earlier studies, many of which, such as Baayen (1989), have been groundbreaking. Rather, the purpose of this research has been to embrace them and to develop them further, by making use of developing technology which may not have been as readily available at the time of the earlier studies, thus providing more sophisticated analyses. The current research has, therefore, examined the studies by Funk (1971), Tottie (1980), Frauenfelder and Schreuder (1992), Schreuder and Baayen (1994), Baayen and Renouf (1996), Kwon (1997), Plag (1999) and Bolozky (1999), but has specifically focused on Baayen and Lieber (1991).

The aims of this research have been various. The first overarching aim was to replicate earlier, established methods of assessing morphological productivity in respect of the negative prefixes, by utilising a larger corpus than has previously been employed. Many previous studies into morphological productivity have adopted the notion, first introduced by Baayen (1989), of the usefulness of hapax legomena in measuring productivity. This study has also endeavoured to re-investigate this phenomenon, and to assess the extent to which hapaxes represent neologisms and, in turn therefore, whether they are indicative of productivity. A number of variables have previously been utilised, either individually or in various combinations, in a variety of methods for measuring productivity, and this research has again compared the different measures in terms of these variables: types, tokens and hapaxes, to
ascertain which variables, or combination of variables, offer the most relevant measure. With the use of the negative prefixes as a case study, a further aim of the research was to reveal their relative productivity and to establish which are currently the most productive. Finally, based on the findings of the above, the study sought to develop more refined methods of measuring morphological productivity.

5.1 Why study morphological negation?

Over the last century, negation as a whole has been well studied by a wide variety of researchers such as Jespersen (1917), Mabbott (1929), Ryle (1929), Klima (1964), Jackendoff (1969), Horn (1989), Croft (1991) and Mazzon (2004), and although all of these touch on morphological negation, none of these specifically concentrate on this form, preferring to focus on sentential negation. The case study presented here, however, has been that of morphological negation; a means of creating negation not previously afforded any great attention in the literature.

Over the past fifty years, very few significant works have focused on morphological negation, with only Zimmer (1964), Funk (1971), Baldi et al (1985), Kwon (1997) and Hamawand (2009) who concentrate their attention on this type of negation. Even amongst these researchers, Funk (1971) and Hamawand (2009) focus on the synonymy of the various negative prefixes involved in morphological negation, and Kwon (1997) examines the negative prefixes diachronically, though he does use a corpus approach. Baldi et al take a psycholinguistic approach, but since their research was carried out in 1985, they do not adopt a corpus-based analysis. Nevertheless, this study remains an insightful and influential piece of work which, as with this thesis, specifically examines the negative prefixes of English. It does not explicitly categorise the prefixes in terms of their degree of productivity, although
the research brings to light several pertinent prefix-specific points, such as the lack of phonological assimilation by the prefix non-. Only Zimmer (1964) investigates morphological negation in terms of the productivity of the negative prefixes of English though, at this time, Zimmer too did not have the facility of electronic corpora, with which to access authentic data in order to measure productivity, despite acknowledging the need for such technology.

As Horn proposes: “In many ways negation is what makes us human, imbuing us with the capacity to deny, to contradict, to misrepresent, to lie, and to convey irony.” (in press:1). Just as productivity is vital for the development and even survival of a language, so negation is also a fundamental concept in human behaviour. Together, the two notions have warranted serious investigation. Although productivity has been given considerable attention over recent years with the advent of corpora, morphological negation has also been in need of similarly in-depth study, which exploits the availability of attested language data. This thesis has succeeded in affording morphological negation the kind of attention its prevalence in everyday language demands.

5.2 Research findings

This thesis aimed to investigate the measures applied by Baayen and his colleagues to determine whether they are indeed the most appropriate for measuring morphological productivity. Unlike earlier studies, it has, however, been able to utilise a much larger corpus than was previously available, along with being able to make use of more sophisticated filtering software. Thus, an extremely large database of negatively-prefixed words was able to be produced which, in turn, allowed a more comprehensive, and therefore, more representative sample to be examined, along
with the use of more contemporary language data than that which was available to earlier studies and, as a result, the previous chapter reported a number of findings.

A variety of previously-tried methods were employed, which utilised a number of variables (types, tokens and hapaxes), but it was found that no single variable used in isolation constitutes a suitable measure of productivity, whereas formulae that involve a combination of variables provide more reliable outcomes. Hapaxes were found to require further analysis in order to allow access to genuine neologisms and, though hapaxes typically include a proportion of neologisms, it was found that this is not generally a large proportion and, more importantly, that this proportion varies very considerably from prefix to prefix. The research has also found that measures which incorporate neologisms as a variable provide more realistic outcomes, and that once neologisms have been calculated, more sophisticated measures of productivity could be applied. The prefix non- was shown to be the most productive of all the five negative prefixes examined, whereas in- was consistently demonstrated as the least productive of the prefixes. Finally, it was found that productivity is affected by word class, and that negative prefixing is a more productive process when nouns are formed, but least productive in terms of the creation of new verbal formations. These findings are explained more closely in the next five sections (5.2.1; 5.2.2; 5.2.3; 5.2.4 and 5.2.5).

### 5.2.1 Which measures of productivity have been replicated?

Having created the database of negatively-prefixed words in Chapter 3, and having examined the historical and semantic factors which may affect the productivity of the negative prefixes, Chapter 4 went on to gather and examine the statistical evidence in terms of replicating a number of previously-tried methods. Initially, type-counting
methods were applied to the negative prefixes and methods such as Aronoff’s (1976) ‘type/potential type’ measure was applied, along with Bolozky’s (1999) ‘dictionary (type) measure. Subsequently, Aronoff’s (1976) and Bolozky’s (1999) measures were rejected as too simplistic, since Aronoff’s measure was based on potential rather than fact, and Bolozky’s measure was judged to be too subjective. In both cases, it was determined that type-counting measures do not indicate the productivity of a process. However, despite the rejection of simple type-counting methods, it was also interesting to note that Hulse (2002) applied a dictionary-based type method (after Bolozky, 1999) to the five negative prefixes and the results are extremely compatible with the findings of the current research, with *non-* classified as the most productive prefix and *in-* as the least productive.

Token-counting methods were next investigated in Chapter 4, and particularly focused on Anshen and Aronoff’s (1988) measure. Although based on early corpora, and despite being compiled from authentic language data, again, the token-counting method was also rejected. Baayen and Lieber (1991) refer to the method as “too simplistic” (807), and again, it is not considered an indication of productivity, since it gives no information as to whether a process is currently used to produce new words. A frequently-used word does not necessarily indicate the productivity of its affix(es); it merely indicates a common word of the language. It was determined that individually both type- and token-counting methods present a purely single-dimensional perspective.

Chapter 4, subsequently therefore, introduced the notion of two-dimensional approaches, which typically engage with and replicate Baayen and Lieber’s (1991) measures. However, it was Aronoff (1976) who initially put forward the notion of a mean type frequency method (Measure A), though Baayen and Lieber (1991)
propose an inverse index of productivity to that which Aronoff proposes. Unlike Aronoff, Baayen and Lieber believe that the lower the mean type frequency, the higher the productivity. The chapter, therefore, went on to investigate this method further in terms of applying it to the negative prefixes, and though the method involves both breadth and depth of coverage, the method can, at best, provide a ‘snapshot’ of the productivity of a process at a given point in time. It allows comparison with other affixes, but only relates to coinages which have occurred up to that point in time, and does not specifically reveal anything about current productivity.

In 1989, Baayen (1989) introduced hapax legomena as a variable in measuring productivity, and, since then, he and his co-authors have taken the logical step from the measure involving types and tokens to the introduction of the hapax/token and the hapax/type methods. Chapter 4 noted that while hapaxes have been widely used in the study of morphological productivity since their introduction, their characteristic properties have been somewhat ignored, rather, they have simply been taken at face value. This was subsequently rectified by examining and discussing the hapaxes more thoroughly in Chapter 4 (and also, therefore, further discussed below). However, in the meantime, the hapax/token and hapax/type methods proposed by Baayen and Lieber (1991) were examined in greater detail.

The hapax/token method (Measure B) was examined first. Baayen and Lieber propose that affixes with low token frequencies are more likely to be involved in the production of new words, and claim that these lower token frequencies indicate more productive affixes (1991:810). Similarly, Chapter 4 also goes on to examine the hapax/type method (Measure C), which proposes that a high hapax/type ratio indicates that an affix is productive. With both of these measures, however, the
emphasis remains on potentiality rather than actuality, and since it had already been determined that the notion of hapaxes required further scrutiny, it was considered necessary to investigate the notion of hapaxes, and the methods involving hapaxes, in greater detail in order to develop more realistic and sophisticated measures.

This research has examined a number of variables in various combinations, based on types, tokens and hapaxes. To some extent, however, it remains a matter of choice as to which variable, or combination of variables, is considered the most relevant to the notion of productivity. Despite the fact that hapaxes require further investigation however, of the five methods explored up to this point, the hapax/type method appeared to come closest to this study’s position on productivity. A list of hapaxes can be seen as a vehicle for accessing neologisms, and in turn, therefore, serves to pinpoint authentic new word formations; in other words hapaxes can help identify genuinely productive processes. Likewise, types refer to the number of individual words which have been formed utilising a particular affix. Thus, hapaxes which have undergone more sophisticated filtering, along with types, ought to provide some information about the productivity of an affix. Token frequency on the other hand, provides a less accurate picture of productivity, since high token frequency tends to distort the picture. A high number of tokens for a particular affix does not reveal the current productivity of a process, rather it comments on the extent to which an affix is incorporated into a commonly-used word of the language.

5.2.2 How useful are hapax legomena in measuring productivity?

The study has re-examined the notion of hapax legomena and has concluded that they are, to a degree, an appropriate way of assessing productivity, since neologisms are more likely to be present in a list of hapaxes than in any other
automatically-identifiable corpus-produced index. However, not all hapaxes are neologisms, and indeed, many are merely infrequently-used words or words which are dying out from the language, and even those that are neologisms do not necessarily have relevance to the particular process that is examined here. Some neologisms, for instance, may be new words by virtue of an alternative word formation rule to the one being analysed. For example, the database of once-occurring negative prefixed words actually included words that had hapax status because of their suffix, not their prefix (for example, *anechoically*, *disentangleable*, *indeterminedness*, *non-parallelistic* and *unbearability*). Moreover, the case study has shown that the proportion of hapaxes that were genuine neologisms differs from negating prefix to negating prefix, which makes global comparisons largely redundant. This is a particularly significant issue, since Baayen’s earlier methods implicitly assume that the neologism-hapax proportion is consistent between all the affixes being compared. In fact, this investigation has shown that this proportion varies considerably from prefix to prefix. Baayen and his collaborators may have subsequently acknowledged that not all hapaxes are neologisms, but this lack of consistency in the hapax-neologism proportion has never previously been recognised. This significant finding is shown in Table 5.1 below, which reproduces Table 4.10 in the previous chapter.

The assumption that neologisms are distributed equally among all the hapaxes is a dangerous one. This current research has clearly shown that proportions differ from process to process. This indicates that the introduction of a manual element is needed in the analysis, and that Baayen and his co-researchers’ findings must be considered as approximations, and must be subject to more fine-grained examination.
Table 5.1. The extent to which hapaxes are neologisms and reflect morphological productivity

<table>
<thead>
<tr>
<th></th>
<th>a-</th>
<th>dis-</th>
<th>in-</th>
<th>non-</th>
<th>un-</th>
<th>Overall percentages per category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misspellings and other errors</td>
<td>0.0% (0)</td>
<td>4.0% (4)</td>
<td>7.0% (13)</td>
<td>1.0% (2)</td>
<td>1.0% (2)</td>
<td>3.0% (21)</td>
</tr>
<tr>
<td>Neologism via suffixation only</td>
<td>2.0% (1)</td>
<td>13.0% (15)</td>
<td>15.0% (28)</td>
<td>1.5% (3)</td>
<td>15.0% (30)</td>
<td>10.0% (77)</td>
</tr>
<tr>
<td>Established word included in OED</td>
<td>78.0% (39)</td>
<td>71.0% (80)</td>
<td>72.0% (136)</td>
<td>12.5% (25)</td>
<td>60.5% (121)</td>
<td>53.0% (401)</td>
</tr>
<tr>
<td>Non-neologism totals</td>
<td>80.0%</td>
<td>88.0%</td>
<td>94.0%</td>
<td>15.0%</td>
<td>76.5%</td>
<td>66.0%</td>
</tr>
<tr>
<td>'GENUINE' NEOLOGISM</td>
<td>20.0% (10)</td>
<td>12.0% (13)</td>
<td>6.0% (12)</td>
<td>85.0% (170)</td>
<td>23.5% (47)</td>
<td>34.0% (252)</td>
</tr>
<tr>
<td>Total no. of hapaxes sampled</td>
<td>(50)</td>
<td>(112)</td>
<td>(189)</td>
<td>(200)</td>
<td>(200)</td>
<td>100.0% (751)</td>
</tr>
</tbody>
</table>

(Figures rounded to the nearest .5 or .0)

5.2.3 Which measure of productivity is most appropriate?

In the first instance, of the previously-tried methods, this research favoured the hapax/token (Measure B) and hapax/type (Measure C) methods, but later concluded that measures which incorporate a type variable were more relevant in measures of productivity, as it was determined that types reflect the actual creation of new complex words using a particular affix. Token frequency, on the other hand, can distort the picture. Frequent use of a word incorporating a specific affix does not necessarily indicate productivity, since words may simply be common everyday words of the language, as was shown in the case of *in-* prefixed words such as *independent, impossible, independence* (the highest frequency *in-* words in the BNC).

However, productivity measures which introduce actual newly-coined words into the equation were shown to provide more sophisticated means of assessing productivity and, once neologisms had been identified, it was possible to apply neologism/token (Measure B1) and neologism/type (Measure C1) methods to
existing productivity formulae, which allowed for these more sophisticated measures to be put into practice. As a consequence, therefore, this research rates the neologism/token and neologism/type measures the highest, which accept the underlying principles of Baayen’s work, but result in more accurate ratings for each prefix. For the reasons given above in terms of token-type choice, and with regard to the importance of neologisms in measuring morphological productivity, Measure C1 (the neologism/type method) is put forward as the most appropriate measure of productivity available, and the measure is replicated (from Table 4.12 in the previous chapter) in Table 5.2 below.

Table 5.2. Measure C1 - productivity via the neologism/type ratio

<table>
<thead>
<tr>
<th></th>
<th>Total No. of hapaxes</th>
<th>Proportion of hapaxes that are genuine neologism</th>
<th>Total No. of types</th>
<th>Level of productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>non-</td>
<td>1,998</td>
<td>0.85</td>
<td>3,784</td>
</tr>
<tr>
<td>2</td>
<td>un-</td>
<td>1,468</td>
<td>0.23</td>
<td>3,836</td>
</tr>
<tr>
<td>3</td>
<td>a-</td>
<td>50</td>
<td>0.20</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>dis-</td>
<td>112</td>
<td>0.12</td>
<td>429</td>
</tr>
<tr>
<td>5</td>
<td>in-</td>
<td>189</td>
<td>0.06</td>
<td>922</td>
</tr>
</tbody>
</table>

Nevertheless, all the variables examined have something to offer to research of this nature, especially if the application of a variety of different measures, which incorporate different combinations of variable, yields results which concur. As Bolozky proposes:

....since no single method constitutes a sure criterion of productivity on its own, several criteria should be applied simultaneously, and a process be characterized as productive when findings based on different methodologies all converge and point to it as such.

(1999:3)
Since all the methods applied which relate to the productivity of the negative prefixes concur, this, therefore, allows the findings of the current research to be presented with even more conviction.

5.2.4 Which prefix is the most productive?

Not only in the neologism/type method described in Table 5.2 above, but across all the measures applied throughout the entire study, non- has consistently emerged as the most productive of the negative prefixes. As described in Chapter 3 (Section 3.5.4), its current productivity may be explained by its purely negative sense, being the only one of the five prefixes to express a strictly contradictory meaning (Funk, 1971:372; Lieber, 2007:114). And even as Zimmer proposed almost fifty years ago in the pre-corpus era, contradictory negation is typically more productive than contrary negation (1964:32).

In contemporary usage, as Mazzon argues, non- is often used “with a pejorative or ironic overtone, or with a euphemistic intention” (2004:112). Likewise, some of its current popularity as a negator may be related to instances where the prefix-free ‘positive’ term has acquired unfavourable or unfortunate connotations, such as non-drinker, non-aerosol, non-aggression, non-smoker or non-blonde, so the negated term in fact carries the positive association and becomes the more frequently used of the two. Certain non- types have also come to be more frequent than their ‘positive’ counterparts; sometimes due to the introduction of a new sense for a word. For example, non-conformist has a frequency of 278 in the database, whereas conformist has a frequency of only 73. Similarly, non-existent has a frequency of 431, compared to a frequency of 96 for existent.
In the same way, *non-* words also help provide an inclusive or exhaustive quality, when used in close proximity to the ‘prefix-free’ positive term. For example:

(i) These new powers of the governing bodies apply to teaching and non-teaching staff, to part-time and full-time and to temporary and permanent staff alike.

(ii) I fell into the conservation trap again with a Sydney chiropractor concerning native, and non-native, birds in Australia.

(Mazzon’s and Marchand’s studies, though thirty-five years apart, respectively describe *non-* as becoming “very productive recently” (Marchand, 1969:180), and as “gaining ground in recent times” (Mazzon, 2004:111). The neutrality of *non-* may serve to increase its contemporary popularity as a negating prefix, whereby it simply means ‘not’ but without any implied criticism, or adverse or negative connotations.

By all measures applied, as the oldest negative prefix of English (see Chapter 3, Section 3.4), *in-* has been shown as the least currently productive of all the prefixes, and indeed, Marchand (1969), Funk (1971), Horn (1989) and Kwon (1997) all describe *in-* as unproductive. Almost a hundred years ago, Jespersen declared, for instance, that the prefix *in-* was unproductive, and suggested that unproductive affixes are learned and stored already incorporated into complex words (1917:140). It has a tendency to combine with common everyday words of the language, which results in greater token frequency, but lesser productivity in terms of the formation of new types.

Chapter 3 (Section 3.5.3) also goes on to propose that *in-* tends towards more negative connotations (cf. *inhuman* and *unhuman*), and is also prone to develop new senses (e.g. the more-commonly used sense of *infinitely*), both of which can affect its
productivity. Marchand claims that “(t)he 20th century does not seem to have made use of the prefix for new formations” (1969:170; also cited in Kwon, 1997:18), and Funk proposes that “syntactic derivation is no longer the domain of in-derivatives in English” (1971:370), while Horn simply describes the in- prefix as “synchronically unproductive” (1989:281). Most of these studies (apart from Kwon, 1997) pre-date the widespread use of electronic corpora, but one of the values of the use of a corpus is in its ability to uphold intuitive opinions with the provision of evidence to verify the theories.

5.2.5 How is productivity affected by word class?

The current study has shown that productivity is sensitive to, and affected by, word class, and it has been demonstrated that negative prefixing is a more productive process when nouns are created and least productive when verbs are formed. Indeed, the same affix can have very different productivity levels depending on the word class of the base to which it is attached.

This thesis argues that a number of factors affect the productivity of individual processes and that, in some respects, assigning a numeric value to the productivity of affixes (as Measures A, B, C, B1 and C1 all did) overlooks the fact that all affixes become more or less productive in certain contexts. For example, word class is clearly a contributory factor in determining which negating prefix is used in a new formation. Some prefixes prefer certain word classes, whereas others tend to avoid particular classes.

From the initial three measures applied (type/token, hapax/token and hapax/type – Measures A, B and C) it has been shown that the productivity levels among different word classes vary markedly between the prefixes, which in turn,
show that productivity is sensitive to word class. The newly-developed measures demonstrate this diversity, with Measure B1 (Table 4.26) ranging from 0.000050 for in- adjectives to 0.116438 for non- adverbs, and for Measure C1 (Table 4.27), the range is from 0.009915 for in- adjectives to 0.495833 for non- nouns. Clearly, word class membership has a significant influence on productivity.

5.2.6 Historical factors and their impact on the productivity of the negative prefixes

Chapter 3 examined how historical factors may affect the productivity of the various negative prefixes, with Baldi et al (1985), for instance, claiming that speakers do utilise etymological information in terms of their productive use of word formation processes.

The time and the circumstances of the introduction of the a- prefix into English have had an effect on the word classes which the prefix favours. For example, from its introduction into English, the a- prefix was typically found in nominal constructions. Likewise, synchronic adjectival formations with a- are still, in reality, formed from nominal bases, rather than merely being the opposite of non-prefixed adjectives (Marchand, 1969:140). Nouns and adjectives are still the most common word classes in terms of a- prefix constructions. In terms of current productivity, however, according to Measures B1 (neologism/token) and C1 (neologism/type), the a- prefix is more likely to produce new adverbial types. The prefixes dis- and in- came into the language incorporated into negative nouns (Kwon, 1997:17-18), and again, this still has an influence on their current predisposition towards nouns. The prefixes dis- and in- tend to combine with more common, everyday words of English, which accounts for their high mean token frequencies. In terms of their current productivity, however, new dis- nouns are rare occurrences and
their neologism/token ratio is the lowest of all the prefixes for all the word classes. When the negative prefix *un*- was introduced into the language, it began to take over from *in*- (Mazzon, 2004:111) and, with a lack of restrictions on the bases it could be attached to, it gained in popularity.

As with the *a*- prefix, *non-* came into English incorporated into nouns, and again, this may account for its current presence and productivity in a considerable number of nominal types, though its current productivity appears to favour adverbial formations. The prefix was originally introduced into English via legal terminology (Mazzon, 2004:111; Kwon, 1997:19) and, though its usage has since spread to other domains, *non-* is still rarely found in literature. The negative prefix *un-* has traditionally been associated with the creation of opposition in adjectives, although according to the neologism/token and neologism/type measures applied, it is currently most productive in nominal constructions.

5.2.7 Semantic factors and their impact on the productivity of the negative prefixes

Semantically, each of the five negating prefixes has its own idiosyncrasies and individual traits. The negative *a*- prefix, for example, is described as forming contradictory opposites, in addition to a small number of privative constructions, and tends to refer to people, as opposed to *in*- which tends to refer to behaviour or actions (Marchand, 1969 and Funk, 1971). Typically, examples such as *amoral*, *asocial* and *atypical* would generally be treated as creating a binary opposition from the positive forms *moral*, *social* and *typical*. Many examples of *a*- prefixed words, as Funk (1971) claims, are from scientific fields, although many examples also have non-scientific applications, for example, *asymmetrical* and *asynchronous*. Mazzon proposes that the *a*- prefix typically refers to a ‘lack of something’, rather than
degrees of a quality of something (2004:111). There are few adverbial a- types and these are not used frequently. Alternatively, in the case of nominal a- types, these are frequent and are used frequently and, thus, constitute robust, common examples. This relates back to the history of the prefix in English, since it was as nouns that it was first introduced in the language, and the prefix has remained a common nominal negator.

Dis- is described by Kwon (1997) as having three discrete but historically connected meanings; the most common of its current senses being one which displays impartiality (disinterested), compared with un- which is more often used attributively (uninterested). Marchand (1969) describes dis- as typically contradictory-forming, such as with the example, disentangle, but with a few contrary instances such as dishonest. Marchand (1969) states that dis- words are common with nouns and adjectives. Although this might well have been the case in the past, the current productivity measures applied in this research do not necessarily support this view. Dis- nominal tokens, for example, have the lowest productivity of all the negative prefixes across all word classes in the B1 Measure (neologism/token ratio), although this is not the lowest productivity level for the C1 Measure (neologism/type ratio). Nominal tokens in dis- tend to be very common, everyday words, rather than literary- or academic-type nouns. Indeed, seven of the highest frequency dis- prefixed words are nouns, which are all still in extremely common currency, despite their longstanding status in the language. Even the most ‘recent’ addition (disappearance (1712)) has been in the language for almost 300 years and most of the other examples have been around for approximately 500 years. Funk (1971) and Kwon (1997) both propose that the semantic distinction between dis- and un- is gradually being lost and dis- is losing out to un-. However, disinterested has a
frequency of 172 in the BNC, whereas uninterested has a frequency of 138. Yet, if the neologism/token and neologism/type productivity measures are examined in terms of dis- adjectives and un- adjectives, un- adjectives score higher in both measures than dis- adjectives (dis- neologism/token - 0.001; neologism/type - 0.031, compared to un- neologism/token - 0.002; neologism/type - 0.087).

Marchand (1969) and Kwon (1997) propose that the prefix in- is a variant of un-, and that like un-, in- can be used to form both contradictory and contrary opposition. The general consensus is that in- almost exclusively combines with adjectives, although there are a few nouns which incorporate the prefix (Marchand, 1969; Mazzon, 2004; Lieber, 2007). The current research finds otherwise. The largest proportion of in- types are, for instance, nouns (38.83%), although in terms of token frequency it is adjectives which make up the highest percentage of in- tokens (64.42%). According to the productivity measures applied by this research, in- is the least productive of the negative prefixes, with few innovations with respect to new types, and likewise, these types also appear to be rarely used in the language with regard to their token frequencies. The prefix non- is described as the most semantically restricted of all the prefixes; in that it creates a purely negative, strictly contradictory and neutral meaning, and is used for classification purposes, as opposed to judgemental descriptions (Hamawand, 2009:131-134). Un- is still described by many of the researchers as the most common and productive of the negative prefixes (Mazzon, 2004:111), but this research shows otherwise, though un- has been demonstrated as particularly common in further-derived forms, such as compounds and novel formations. However, by all the measures applied by this research, non- has been demonstrated as the most productive, despite Funk’s view
that it is rarely used in everyday language (1971:372), and Lieber’s opinion that *non-* is the most semantically restricted of the negative prefixes (2007:114).

Hamawand’s taxonomy (referred to in Chapter 3, Section 3.5.6 – Table 3.4), though comprehensive, exclusively focuses on semantics and does not refer to productivity in any way. Nevertheless, it is apparent that all of these factors may have, to some extent, affected the productivity of the negative prefixes diachronically, from their introduction into the language up to more recent times. However, it is not necessarily the case that these factors continue to affect productivity. Clearly, some factors continue to restrict productivity, whereas others now have no bearing on current trends in language, or on the productivity of the negative prefixes, and indeed other affixes.

5.3 Have the aims of this research been fulfilled?

Generally speaking, this research agrees with Baayen’s innovative quantitative measures and, to a large extent, with his hapax-based approach. It has, however, fine-tuned the methods by using a larger corpus than was available to Baayen in the early 1990s, and has developed a more discerning approach to the use of hapaxes, in an attempt to obtain optimum results.

Central to this thesis has been an examination of the relationship between hapaxes and neologisms. Rather than being single-frequency words which simply represent neologisms, this study has found that hapaxes can be categorised into distinct groups, and that the group which comprises genuine neologisms only constitutes approximately one third of all hapaxes, whereas over half of all hapaxes are, in fact, established words of the language. More significantly, this research has also determined that the proportion of neologisms in a list of hapaxes varies
considerably from affix to affix which, in the case of the negative prefixes, ranges from 6% (for *in-* ) to 85% (for *non-* ) (Table 5.1).

The research has also examined types and tokens to determine which are the more reliable in terms of a measure of productivity, and has found that while tokens indicate frequency of use, it is types which best reflect the creation of a new word form and, thus, actual productivity. In addition to their individual examination, types and tokens have been investigated in relation to each other, by means of type/token ratios. Hapax/token and hapax/type relationships have also been analysed, along with observations which relate to frequencies greater than single occurrences.

Likewise, the study has undertaken examination of the distribution of the negatively-prefixed words of the database in terms of word class. Again, with regard to word class, this research has examined types and tokens both individually and in combined relationships, as well as comparing these distributions to that of the word class distributions for the *BNC* as a whole.

### 5.4 Limitations, recommendations and future research

While the availability of corpora has undoubtedly allowed empirical studies to take a massive step forward, this research has also identified the need for more sophisticated analysis than simply the use of quantitative measures alone. Corpora allow vast quantities of authentic language data to be examined both in terms of quantitative analysis and by means of the provision of contextual information. This research has, to some extent, adopted Van Marle’s (1992) opinion, when he states: “I very much doubt whether Baayen’s approach to equate hapaxes with newly coined formations (and, consequently, to consider them indicators of productivity) is generally tenable” (157). As with this current study, Van Marle does not imply that
Baayen’s hapax-based approach lacks in merit. Indeed, Van Marle, while disagreeing with Baayen on a number of points, also expresses the view that:

The overall importance of Baayen’s work is undisputable and it seems unquestionable that his work will greatly deepen our understanding of some of the basic properties of the phenomenon of morphological productivity.

(1992:152)

While the main focus of this research has been on a corpus-based approach to morphological productivity, it has also employed a qualitative strand by the use of the *OED* to qualify the presence of neologisms within the hapax lists and, in addition, has also replicated a dictionary-based measure proposed by Bolozky (1999), which has similarly supported the other findings of this study. Therefore, the introduction of a qualitative element, and indeed a manual element, to this research, has facilitated a more in-depth investigation of the corpus-generated data in order to quantify the neologisms contained in the lists of hapaxes, and it was in this regard that the *OED* was employed. As Bolozky (1999) and Plag (1999) have separately proposed, whilst text-based corpora methods have proved to be extremely useful in measuring productivity, the use of a combination of methods, including dictionary approaches, can also serve to validate and support findings. Bolozky, for example, argues that:

....comparing findings arrived at through different procedures is the safest evaluation measure. The higher the correlation between findings of different measurements conducted independently of each other, the higher the degree of validity.

(1999:192)

Plag, too, believes that dictionaries that contain historical data, such as the *OED*, whilst they have certain drawbacks in terms of the production of a database, are
useful with regard to supporting information (1999:102). Plag believes that Baayen’s corpus-based measures are entirely valid, but proposes that these quantitative measures can, and should, be complemented by a qualitative element to reinforce the findings. (1999:244). Thus, the various elements used in this research have served to justify the methods and support the findings.

5.4.1 Limitations of this research

At the start of this research, the decision was made to use the one-hundred-million-word British National Corpus, as the most comprehensive and readily-available at the time. Even over the course of this study, larger corpora have become available, but the decision had already been made to utilise the *BNC* for a variety of reasons. The texts of the *BNC* are taken from a vast array of “different kinds of speech and writing (e.g. conversation, novels, news reporting) broadly in accordance with their representation in everyday language use.” (Leech et al, 2001:1). Although the *BNC* covers speech as well as writing, which some other corpora do not, the *BNC* is made up of 90% written texts and 10% spoken. This may be considered an issue, since it could be argued that the spoken word takes primacy over written texts in terms of human communication. For example, all languages have a spoken or signed form, whereas not all languages have a written form. However, apart from purely spoken corpora which tend to be smaller overall due to the skill needed and the time-consuming nature of transcribing spoken language into a usable written format, the *BNC* is probably remains one of the most comprehensive and representative bodies of language data.

The database on which this study is grounded could be criticised on a number of levels. The initial production of the database made use of software which was
readily able to identify words which begin with ‘dis-’, ‘in-’, ‘non-’ and ‘un-’. Identifying words which begin with ‘a-’, however, was not so straightforward, since every word in the BNC which began with ‘a-’ was extracted, and manual editing had to be employed to specifically identify negatively a- prefixed words. Therefore, whereas the Sketch Engine software was used for the other prefixes, a- prefixed words were extracted by other means. Not only did the ‘a-’ words have to be manually edited to determine the a- prefixed words, but the list also then had to be screened to remove a- prefixed words which were not negatively a- prefixed words, such as alike, alive, asleep, etc., and again, this process had to be conducted manually. Furthermore, in the case of negatively a- prefixed words, decisions also had to made based on whether the prefix was the variant introduced into English already incorporated into words, for example, anonymous /ənəˈmɔːrəs/ or amnesia /əˈmɛsnɪə/, where the base word does not exist in its own right in English (*onymous or *mnesia), as opposed to the living English prefix /əl/, as in atypical. Although generally the distinction is quite clear, a very few examples were determined as one type or the other based on personal intuition, thus introducing a subjective element. For instance, in acephalous, anaesthetic, agnostic and apathy, English base words are present but these are either rare words in English or they are used in a slightly different sense. The manual element in editing the corpus-generated list does not, however, detract from the usefulness of corpora-gathered data.

Likewise, although the shortcut software allowed words beginning with ‘dis-’, ‘in-’, ‘non-’ and ‘un-’ to be identified easily, it still could not discern between words which simply began with this combination of letters, such as disaster, inferior, nonsense and university, and those which were negatively-prefixed words. Again, the database had to be edited manually. To avoid potential criticism of this manual
editing, criteria were determined and strictly adhered to in terms of the editing. In cases which were ambiguous, contextual evidence was also examined. For instance, as discussed in Chapter 3, Section 3.3.2, in the case of the word disputative, it is unclear as to whether this is disputative in the sense of ‘inclined to argument or dispute’ or whether this is disputative meaning ‘not putative’ or ‘not commonly regarded as’. In case such as this, the contextual information was consulted. The above demonstrates that whilst electronic corpora allow for greater objectivity by means of empirical analysis, some degree of subjectivity is inevitable, though this does not detract from the validity of the corpus approach, since, above all, it readily and transparently facilitates replication.

5.4.2 Recommendations and future research

This study has focused on productivity and particularly that of the negative prefixes, and has determined that hapax legomena are an appropriate means of accessing neologisms, in that a list of hapaxes will inevitably include a percentage of neologisms, though this percentage varies from prefix to prefix. It would be interesting to discover whether measures based on a data collection system which specifically compiles a straightforward list of neologisms would yield the same or similar results. Plag (2003), for example, proposes the use of the CD-ROM version of the OED, which can be utilised to search for neologisms for a specific period of time (53). He acknowledges that “(t)he power of the OED as a tool for measuring productivity should not however be overestimated”, since new words are not guaranteed to be included in the OED (53). One of the advantages of the OED search tool, however, is that the OED includes comprehensive information on a particular word and “allows us to trace neologisms for any given period of time.” (Plag,
Nonetheless, a tool which directly identifies neologisms should be seen as a significant benefit in measuring productivity.

The examination of the negative prefixes in terms of their domain preferences and their written/spoken distribution has only been an indicative measure in this study, though as with word class, it is apparent that domain also has an effect on the choice of one prefix over another. Certainly this is an area which has been overlooked previously and may well benefit from further investigation. As discussed in earlier sections, this research has considered both written and spoken language, as the BNC includes both genres, though there is an obvious bias towards written language. The study has also examined the domains which the different negative prefixes favour, which again incorporates both written and spoken data. Whilst the BNC is a representative sample in that it incorporates data from a variety of domains, nevertheless, the bias is still towards written text (Plag, 2003:206). It may be interesting, however, to compare written and spoken language on a more equal footing in terms of productive processes and negative prefix distribution. A corpus with an equal split between written and spoken language, or alternatively two separate but comparable written and spoken corpora could be used to compare the two types of discourse. Furthermore, as the choice of one prefix over another also appears to be affected by domain, the opportunity arises to assess productivity across different domains. In which genre is productivity greatest? Do different affixes have higher productivity levels in different domains? Are neologisms more common in speech or writing? Just as this study has shown that productivity levels vary according to word class, so other factors could be investigated in order to develop a fuller picture of how new complex words enter the language.
Overwhelmingly, this research has identified the prefix *non-* as the most productive of the negative prefixes. This finding is, of course, based on a corpus which was compiled in the 1990s and is a fixed and static corpus. As proposed by Sinclair (1991), a monitor or dynamic corpus, which constantly changes and develops, could help future studies make even more robust judgements about productivity, and indeed about changing productivity (25). A dynamic corpus of this kind would provide the facility to “monitor changing patterns of usage over time” (Kennedy, 1998:60). According to the application of Bolozky’s (1999) dictionary-based measure to the negative prefixes (Table 4.1, Section 4.1.1.2), and also the evidence of the current study, the trend over the past thirty-five years appears to show this increasing productivity of *non-*; but it would be interesting to see if this trend persists, is perhaps slowing down, or whether perhaps the other negative prefixes are now ‘fighting back’. For example, according to Measure C1 (Table 4.27, Section 5.3.6), negatively-prefixed nouns are the most productive by the neologism/type measure, with nouns prefixed by *un-* constituting the highest proportion of all the *un-* types. This supports Horn’s (2002) opinion of the current increase in the productivity of novel *un-* nouns, and could possibly also indicate a changing trend towards an increase in the productivity of *un-*.

Finally, adverbs prefixed with *non-* have been shown as the most productive of all the word classes. Potentially, this is a surprising result. Typically, the prefix *non-* is associated with contradictory or non-gradable opposition. Adverbs, on the other hand, typically ‘modify’, rather than create a direct binary opposite. Although it is outside the immediate scope of this research, it would be intriguing to investigate whether *non-* is drifting semantically towards more contrary opposition, as opposed to its more common contradictory status. Similarly, though adverbs
would not traditionally be assumed to be the most productive of word classes, it would also be interesting to examine whether the current productivity of non- adverbs is the result of the existing large number of non- adjectives from which most adverbs are derived.

5.5 Chapter and thesis summary

This thesis serves to support the findings of Baayen and his colleagues in terms of morphological productivity but has sought to further develop their methodologies in order to fine-tune the measures. It should, therefore, be viewed both as a compliment and a complement to Baayen’s innovative early corpus studies, and as a precursor to further data-based analyses. This research has had the advantage of being able to utilise a much larger corpus than Baayen had at his disposal in the early 1990s, so that, although his measures have been replicated, the database itself used in this research has been much larger and, therefore, the findings are more robust and more representative of language as a whole.

This study does not detract from Baayen and his co-researcher’s work, but adds to its validity. This thesis has found that, as Baayen proposed in 1989, hapax legomena are a useful and appropriate means to gain access to new words of the language. Hapaxes are not equivalent to neologisms, but neologisms will be found within a list of hapaxes, and this research has determined that the proportion of hapaxes which are neologisms varies considerably from prefix to prefix, which makes global comparisons between prefixes largely redundant. For words negatively-prefixed with in- for example, the neologisms within the list of hapaxes represented only 6.0% of the total number of hapaxes, whereas for non- prefixed words, the neologisms accounted for 85.0% of the hapaxes. This is based of
self-designed criteria based on *OED* inclusion. The fact that neologisms are not equally distributed among the hapaxes but vary from process to process demonstrates that Baayen’s methods can only be viewed as approximations, but that with the introduction of a manual element to the procedure and a more refined use of hapaxes, a more sophisticated measure can be applied with more definitive findings.

Measures have been employed in this study which take into account a number of variables, such as types, tokens, hapaxes and neologisms, but all have pointed to the prefix *non-* as the most productive negative prefix in English. In addition, contextual, semantic, historical and domain information have also been employed in this thesis, along with an examination of the distribution of the prefixes in terms of word class. All of these facets are inter-related, and have all been valuable in assessing how the individual prefixes will be used in the future, and in how the phenomenon of morphological productivity can best be evaluated.
BIBLIOGRAPHY


Chomsky, N. (1962). Paper given at the University of Texas 1958, 3rd Texas Conference on Problems of Linguistic Analysis in English, Austin, University of Texas.


Websites

BBC. http://www.bbc.co.uk/worldservice/learningenglish/grammar/learnit/learnitv45.


http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html


The National Curriculum. www.nc.uk.net (accessed 02/02/10).


Qualifications and Curriculum Development Authority.


